AD-A119 915 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
STATISTICAL MODELLING OF BIVARIATE DATA.(U)
AUG 82 T J WOODFIELD
UNCLASSIFIED TR-8-7 DAAG29-80-C-0070
ARO-16992.13-MA NL

12

STATISTICS

TEXAS A&M UNIVERSITY

ARO 16992.13-MA

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843-3143

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

STATISTICAL MODELING OF BIVARIATE DATA

Terry Joe Woodfield

Institute of Statistics, Texas A&M University

Technical Report No. B-7

August, 1982

Texas A&M Research Foundation
Project No. 4226

"Robust Statistical Data Analysis and Modeling"

Sponsored by the U. S. Army Research Office
Grant DAAG29-80-C-0070

Professor Emanuel Parzen, Principal Investigator

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report B-7	2. GOVT ACCESSION NO. A119 915	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Statistical Modeling of Bivariate Data		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Terry Joe Woodfield		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0070
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE August 1982
		13. NUMBER OF PAGES 263 + vi
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

joint density-quantile function, dependence-density, non-parametric bivariate density estimation, entropy, exponential models for densities

20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A technique for modeling bivariate data that is based on the theory of orthogonal expansions in a separable Hilbert space is examined. A new nonparametric density estimation procedure is developed using an information criterion and is shown to be equivalent to least squares estimation of a density when the criterion function is computed with respect to the empirical distribution function. Computer programs are presented that implement the procedure for the univariate and bivariate cases. Examples utilizing these programs are given and comparisons made to existing density estimation techniques.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

PROLOGUE

by

Emanuel Parzen

Institute of Statistics
Texas A&M University

This Ph.D. thesis is an important contribution to a new dimension to statistical reasoning for which I propose the name FUN STAT (because it is fun; functional (useful); based on functional analysis; estimates functions; and graphs functions). FUN.STAT has three important components: quantile and density-quantile signatures of populations, entropy and information measures, and functional inference.

The joint density quantile function of (X,Y) where X and Y are jointly continuous random variables can be represented

$$fQ_{X,Y}(u_1,u_2) = f_{X,Y}(Q_X(u_1), Q_Y(u_2)) = fQ_X(u_1)fQ_Y(u_2)d(u_1,u_2)$$

in terms of the marginal density-quantile functions $fQ_X(u)$, $fQ_Y(u)$, and the dependence density $d(u_1,u_2)$. How these three functions can be semi-automatically estimated, by autoregressive or exponential model estimators with maximum entropy properties, is investigated in this thesis. The results provide important and useful procedures for nonparametric bivariate density estimation. The thesis discusses estimators of the entropy $H(d)$ of $d(u_1,u_2)$, which seem to me to be important because they can be applied to provide a useful quality-index for projection-pursuit data analysis methods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 The Problem

Much of statistical analysis revolves around the interrelationship between random variables. One may explore cause and effect relationships, investigate the covariance structure of a collection of random variables, or attempt to discover the underlying probability mechanism that produces a vector of random variables. The areas of regression and correlation analysis, analysis of variance, categorical data analysis, and the general area of multivariate analysis attempt to confront some of the relevant problems in dealing with relationships among random variables. Mathematical tools from probability theory and the theory of vector spaces assist in analyzing the abstract problem, but one must also overcome computational difficulties that arise from examining discrete observations from a continuous multivariate distribution. The esoteric nature of statistical analysis results from the wide range of mathematical and computational tools that must be employed in solving general data analytic problems. In this work we attempt to consolidate a variety of such tools to provide a solid base from which to attack the general problem of multivariate data analysis. We have chosen bivariate data modeling as the logical starting point, and that is the primary subject of this thesis; however, multivariate generalizations will be

---

This dissertation will follow the format for the Journal of the American Statistical Association.

suggested whenever appropriate.

In its most general form, the problem is to infer from a bivariate random sample $\{(X_i,Y_i),\ i=1,\ldots,n\}$ the nature of the joint cumulative distribution function (c.d.f.)

$$F_{X,Y}(x,y) = P(X\le x, Y\le y)$$

and the marginal cumulative distribution functions

$$F_X(x) = P(X\le x),\ F_Y(y) = P(Y\le y).$$

Knowledge of these entities will answer most of the questions posed in regression and correlation analysis, but more generally, information will be provided about the dependence structure between the random variables X and Y. The theory of parametric inference is based on assumptions concerning these functions, but one is still faced with the problem of testing these assumptions. We will emphasize the problem of determining the dependence structure between two random variables, but our approach will lead to solutions to more general problems of bivariate data analysis.

A specific application will be to provide techniques for testing the null hypothesis

$$H_o:\ X \text{ and } Y \text{ are independent}$$

against some suitable alternative. Many useful techniques already

exist for handling this problem, but often such techniques carry restrictive assumptions or demand too much computational complexity. We will propose a general technique carrying few restrictions that is computationally manageable and that suggests applications to other areas of bivariate data modeling.

1.2 Survey of the Literature

A wide variety of sources exist from which to extract useful information for attacking the problem of bivariate data analysis. For general nonparametric measures of association, Lehmann (1966), Blomqvist (1950), Blum, Kiefer, and Rosenblatt (1961), and Hoeffding (1948) provide useful fundamental information. Puri, Sen, and Gokhale (1970) and Parzen (1977) contain useful discussions of independence tests in a multivariate setting. Classical normal theory is exemplified in Morrison (1976) with more theoretical results appearing in Kshirsagar (1972) and Rao (1973).

The approach we employ considers function approximation in a nonparametric setting using information measures. References on nonparametric density estimation are provided by Rosenblatt (1956), Parzen (1962), Cacoullos (1966), Loftsgaarden and Quesenberry (1965), Kronmal and Tartar (1968), Tartar and Kronmal (1970,1976), Crain (1974), Carmichael (1976), and Good and Gaskins (1980). General expository and bibliographic sources are provided by Tapia and Thompson (1978), Bean and Tsokos (1980), Wertz (1978), and Silverman (1980). For the mathematical theory of function approximations,

Lanczos (1956), Rainville (1960), Davis (1975), and Powell (1981) are useful texts. Loève (1977), Hewitt and Stromberg (1965), and Royden (1968) contain some results from functional analysis that may also be applied to this problem in a measure space setting. Parzen (1959,1961) also contains some useful function space results in a statistical setting. Shannon (1948) and Kullback (1978) are the fundamental references for information theory.

The motivation for much of this research is provided by Parzen (1977,1979), Kimeldorf and Sampson (1975b), Crain (1974), and Tartar and Kronmal (1970). The quantile domain approach to statistical data modeling found in Parzen (1979b) provides some useful solutions that can be extended to bivariate data analysis, with Kimeldorf and Sampson (1975b) providing some useful bivariate theory to apply to the problem. The orthogonal expansion technique as a method of nonparametric density estimation seems to be the best suited for multivariate extension of univariate methods. The ideas of Crain (1974) and Tartar and Kronmal (1970) motivate the development of a modification of their techniques based on a general regression framework using information theoretic notions.

Scott, et al. (1978), employ the bivariate kernel method to a set of coronary heart disease data. Tartar and Silvers (1975) apply orthogonal expansion techniques to the problem of bivariate Gaussian mixture decompositions. These applications suggest a need for a more objective and less cumbersome approach to the problem of diagnosing the shape of a bivariate density, which motivates the applications considered in the present work.

Finally, Csörgö and Révész (1981), Serfling (1980), and Randles and Wolfe (1979) provide comprehensive accounts of the relevant asymptotic distribution theory for many of the fundamental statistics.

2. MATHEMATICAL AND STATISTICAL FUNDAMENTALS

2.1 Introduction

The approach to bivariate data modeling that we will develop in Chapter 4 is motivated by three concepts: 1) the use of quantile based data analytic tools; 2) the use of information and entropy criterion functions; and 3) the application of some powerful results from approximation theory. This chapter provides expository information on these subjects along with a few remarks and observations that may not be found in the literature. An additional section is included describing some elements of stochastic processes and complex regression applicable to the models employed in Chapter 4. We have assumed knowledge of basic mathematical statistics similar to that found in Rao (1973).

2.2 Uniform Representations and the Probability Integral Transform

In this section we introduce concepts of probability modeling set in the quantile domain, i.e., a domain of consideration in which the quantile function is the fundamental entity. The foundation of much of the theory will be directly or indirectly related to the probability integral transform which makes the quantile approach so appealing.

Let X be a random variable (r.v.) with c.d.f. F and probability density function (p.d.f.) f. Define the quantile function Q(u) of X

by

$$Q(u) = F^{-1}(u) = \inf\{x: F(x) \geq u\}, 0 \leq u \leq 1. \qquad (2.2.1)$$

When two or more r.v.'s are considered, one affixes a subscript and denotes $Q(u)$ as the quantile function of X, etc. This definition of a quantile function yields a result known as the correspondence identity, namely

$$F(x) \geq u \text{ iff } Q(u) \leq x. \qquad (2.2.2)$$

(The expression iff is the commonly used mathematical abbreviation for "if and only if".) When F is continuous, one has the inverse identity

$$FQ(u) = F(Q(u)) = u. \qquad (2.2.3)$$

Differentiating the inverse identity, one obtains the reciprocal identity

$$fQ(u)q(u) = 1. \qquad (2.2.4)$$

The notation $fQ(u)$ refers to the density-quantile function defined to be the composite function $f(Q(u))$. One also has the quantile-density function $q(u)$ defined to be the derivative of the quantile function. Another useful function is the negative of the derivative of $fQ(u)$, often called the score function, given by

$$J(u) = -f'Q(u)q(u). \tag{2.2.5}$$

The score function is usually written

$$J(u) = -f'Q(u)/fQ(u). \tag{2.2.6}$$

Randles and Wolfe (1979b) call $J(u)$ the optimal score function. One application for $J(u)$ involves the concept of information. Consider a form of Fisher's information,

$$I(f) = \int_{-\infty}^{\infty}[\frac{\partial}{\partial x}\log f(x)]^2 f(x)\,dx$$

$$= \int_{-\infty}^{\infty}\frac{|f'(x)|^2}{f(x)}dx = \int_0^1 |J(u)|^2 du. \tag{2.2.7}$$

Thus, this information measure requires only knowledge of the score function. For Shannon entropy, the density-quantile function is the fundamental object, namely

$$H(f) = \int_{-\infty}^{\infty}-[\log f(x)]f(x)\,dx$$

$$= \int_0^1 -\log fQ(u)\,du. \tag{2.2.8}$$

With the quantile building blocks considered above, one may now state two fundamental theorems that will be exploited later.

Theorem 2.2.1 Let the r.v. U be distributed uniformly on the

interval [0,1], and let F be a c.d.f. Define the r.v. X by X=Q(U) where Q is the quantile function associated with F. Then the c.d.f. of X is F.

Proof: $P(X \le x) = P[Q(U) \le x] = P[U \le F(x)] = F(x)$.■

Theorem 2.2.2 Let X be a r.v. with continuous c.d.f. F. Then U=F(X) is a uniform r.v. on the interval [0,1] (In the proof, Q(u) is the quantile function of X.)

Proof: $P(U \ge u) = P[F(X) \ge u] = P[X \ge Q(u)] = P[X > Q(u)]$

$= 1 - P[X \le Q(u)] = 1 - FQ(u) = 1 - u$.■

Details using some of the aforementioned identities are omitted from the above proofs but may easily be supplied by the reader. The transformation U=F(X) is called the probability integral transform and is very useful in attacking general problems in such a way that only uniform distributions need be considered. The probability integral transform also reduces the general simulation problem to one of simulating uniform [0,1] random variables. One calls U the uniform representation of X. This terminology will become more meaningful in the bivariate case. In the univariate case, any continuous random variable has the same uniform representation. The usefulness occurs when results are invariant to the probability integral transform.

Moments may be considered in the quantile domain by applying some of the results obtained above. Observe,

$$\mu = E(X) = E[Q(U)] = \int_0^1 Q(u)\, du \qquad (2.2.9)$$

from Theorem 2.2.1, and

$$\sigma^2 = Var(X) = \int_0^1 [Q(u)-\mu]^2 du. \qquad (2.2.10)$$

Another application involves transformations of the form $Y=g(X)$ where X has a known distribution. A common transformation is the location-scale transformation

$$Y = \mu + \sigma X. \qquad (2.2.11)$$

One is given $Q_X(u)$ and wishes to obtain $Q_Y(u)$. *Observe,*

$$F_Y(y) = P(Y \le y) = P(\mu+\sigma X \le y) = P[X \le (y-\mu)/\sigma] = F_X[(y-\mu)/\sigma].$$

$$F_Y(y) \ge u \text{ iff } Q_Y(u) \le y$$

is equivalent to

$$F_X[(y-\mu)/\sigma] \ge u \text{ iff } Q_X(u) \le (y-\mu)/\sigma.$$

Furthermore,

$$Q_X(u) \le (y-\mu)/\sigma \text{ iff } \mu+\sigma Q_X(u) \le y.$$

Hence, it follows that

$$Q_Y(u) = \mu + \sigma Q_X(u). \qquad (2.2.12)$$

One may seek similar results for general transformations $Y=g(X)$. Let g be a strictly increasing function. Then,

$$F_Y(y)=P[Y\le y]=P[g(X)\le y]=P[X\le g^{-1}(y)]=F_X[g^{-1}(y)].$$

Again, the correspondence identity for Y is equivalent to

$$F_X[g^{-1}(y)] \ge u \text{ iff } Q_X(u) \le g^{-1}(y)$$

and

$$Q_X(u) \le g^{-1}(y) \text{ iff } g[Q_X(u)] \le y.$$

Thus,

$$Q_Y(u) = g[Q_X(u)]. \qquad (2.2.13)$$

Now, suppose g is strictly decreasing. It follows readily that

$$F_Y(y) = 1 - F_X[g^{-1}(y)].$$

The correspondence identity for Y is equivalent to

$$F_X[g^{-1}(y)] \leq 1-u \text{ iff } Q_X(1-u) \geq g^{-1}(y)$$

and

$$Q_X(1-u) \geq g^{-1}(y) \text{ iff } g[Q_X(1-u)] \leq y.$$

Thus,

$$Q_Y(u) = g[Q_X(1-u)]. \qquad (2.2.14)$$

Parzen (1979b) considers the general problem of transformations of random variables to specified distributions (such as normal) in light of the above results.

Some useful extensions to these concepts may be applied to goodness-of-fit (g.o.f.) procedures. If one defines $D(u)=FQ(u)$ and $d(u)=D'(u)$, these represent the c.d.f. and p.d.f. respectively of a uniform (0,1) random variable. For a null hypothesis H_o: $fQ(u)=f_oQ_o(u)$, Parzen (1979b) calls

$$d(u) = f_oQ_o(u)\,q(u) / \int_0^1 f_oQ_o(u)\,q(u)\,du \qquad (2.2.15)$$

the <u>f_oQ_o transformation density</u> which is a uniform (0,1) density under the null hypothesis. The statistical applications of (2.2.15) will be considered in section 3.5 in the discussion of autoregressive density estimators. Parzen also discusses <u>tail-exponents</u> as a means of classifying distributions based on density-quantile representations.

The reader may consult Parzen (1979b) for more extensive results in the univariate theory.

Some of the above univariate concepts extend readily to the bivariate case. Let X and Y be continuous r.v.'s with joint c.d.f. $F_{X,Y}$ and marginals F_X and F_Y. Let the respective quantile functions be denoted by Q_X and Q_Y. Define the dependence distribution function $D(u_1,u_2)$ by

$$D(u_1,u_2) = F_{X,Y}(Q_X(u_1),Q_Y(u_2)),\ 0\le u_1,u_2\le 1. \qquad (2.2.16)$$

Parzen (1977) calls $D(u_1,u_2)$ the regression distribution function, while Kimeldorf and Sampson (1975b) call it the uniform representation of $F_{X,Y}(x,y)$. The dependence density $d(u_1,u_2)$ is given by

$$d(u_1,u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} D(u_1,u_2) = \frac{f_{X,Y}(Q_X(u_1),\ Q_Y(u_2))}{f_X(Q_X(u_1))f_Y(Q_Y(u_2))}. \qquad (2.2.17)$$

Note that while the univariate representations of the above objects are related to the uniform (0,1) distribution and hence have extensions to goodness-of-fit procedures, the dependence distribution function and the dependence density have the added bivariate role of detecting independence between two random variables, justifying the name we have given them. Furthermore, they correspond to bivariate r.v.'s distributed uniformly on the unit square only when X and Y are independent, that is, $D(u_1,u_2)=u_1u_2$ and $d(u_1,u_2)=1$ if and only if X and Y are independent. More general bivariate uniform distributions

are considered by Kimeldorf and Sampson (1975b).

Since the bivariate normal distribution is usually the "null hypothesis" distribution one may be interested in the shapes of the various functions of interest. Figure 1 depicts a bivariate standard normal p.d.f. with the correlation coefficient equal to zero, while Figure 2 shows the linear concentration of the probability mass when the correlation coefficient is equal to .9. Figure 3 shows the dependence density corresponding to Figure 2, while for the independence case, the dependence density is a flat surface identically equal to one. Figures 4 and 5 show the bivariate density-quantile functions corresponding to Figures 1 and 2. Figures 3 through 5 have not been observed in the literature although they contribute insight into the relationships between the various functions of interest.

One may establish an equivalence relation based on the above uniform representation. Two bivariate distribution functions $F_{X,Y}$ and $G_{X,Y}$ are said to be equivalent (written $F_{X,Y} \sim G_{X,Y}$) if $D_F = D_G$ where the subscript refers to the distribution for which the uniform representation is defined. Thus, all bivariate distribution functions of independent random variables are equivalent in this sense. The following Theorem allows one to apply this concept to generating new distributions with arbitrary prescribed marginals.

Theorem 2.2.3 (Kimeldorf and Sampson, 1975b) Let $F_{X,Y}$, $G_{X,Y}$ be bivariate distribution functions with associated marginals F_X, F_Y, and G_X, G_Y and corresponding quantile functions F_X^{-1}, F_Y^{-1} and G_X^{-1}, G_Y^{-1}.

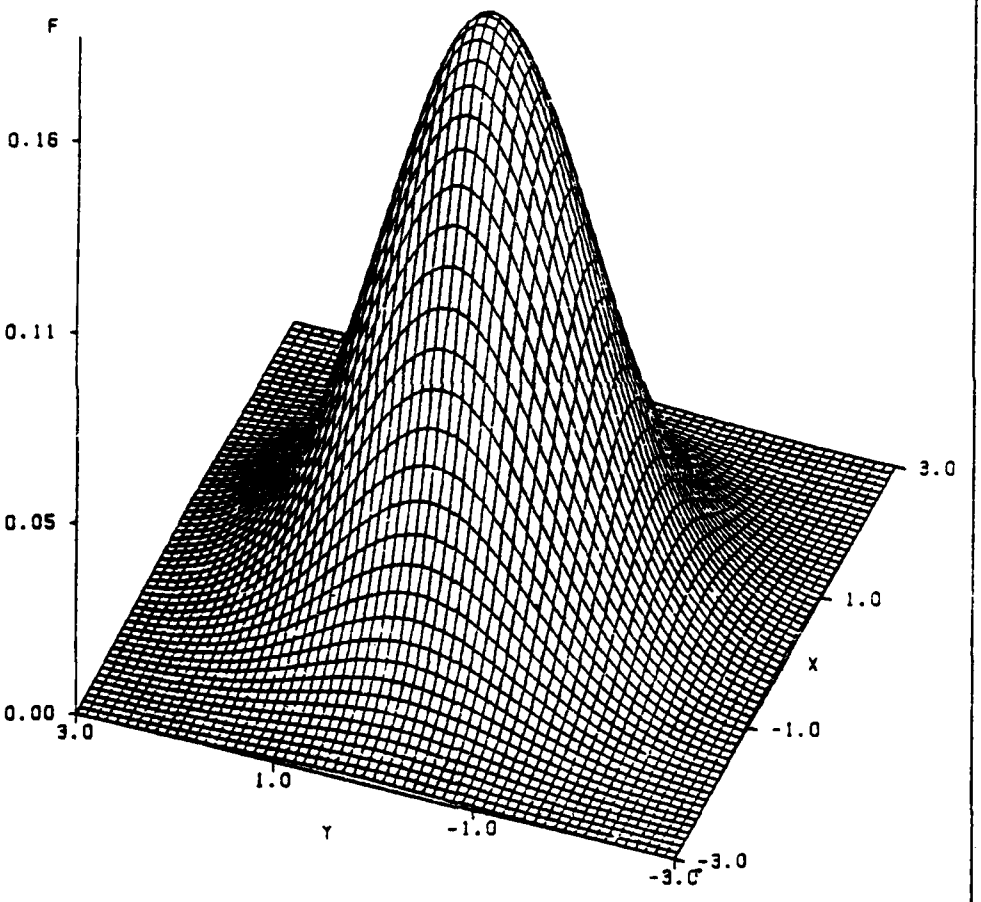


Figure 1. Bivariate Standard Normal P.D.F. for Rho=0

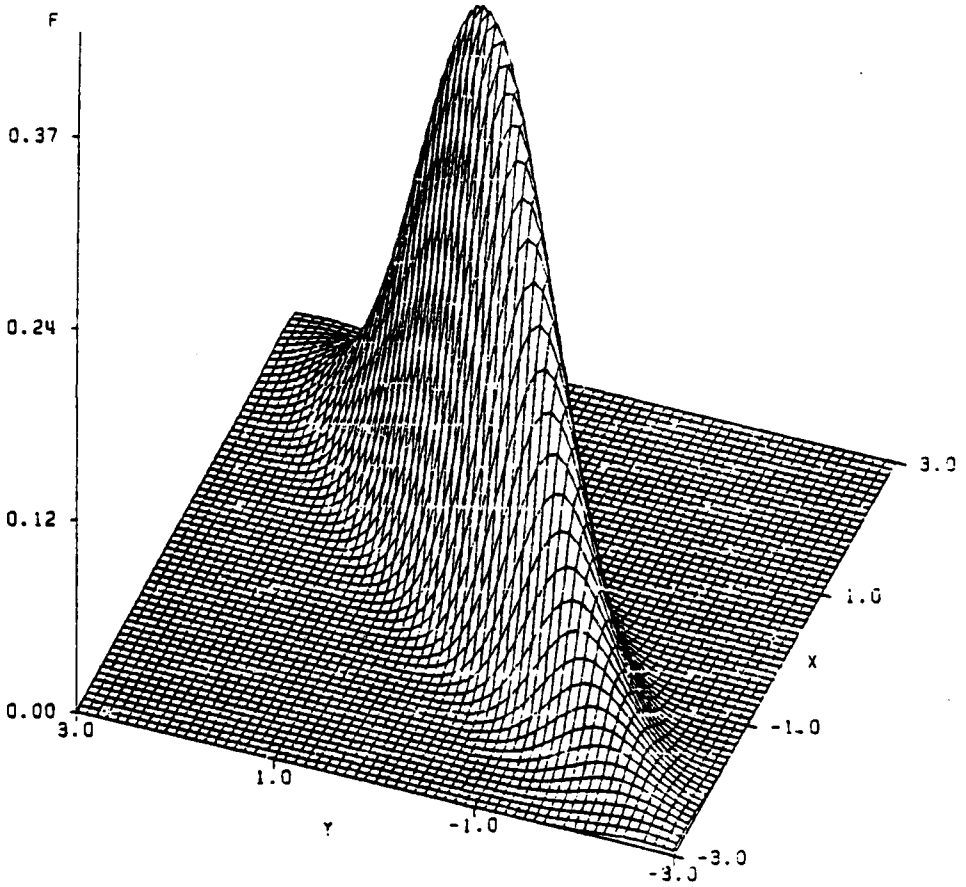


Figure 2. Bivariate Standard Normal P.D.F. for Rho=0.9

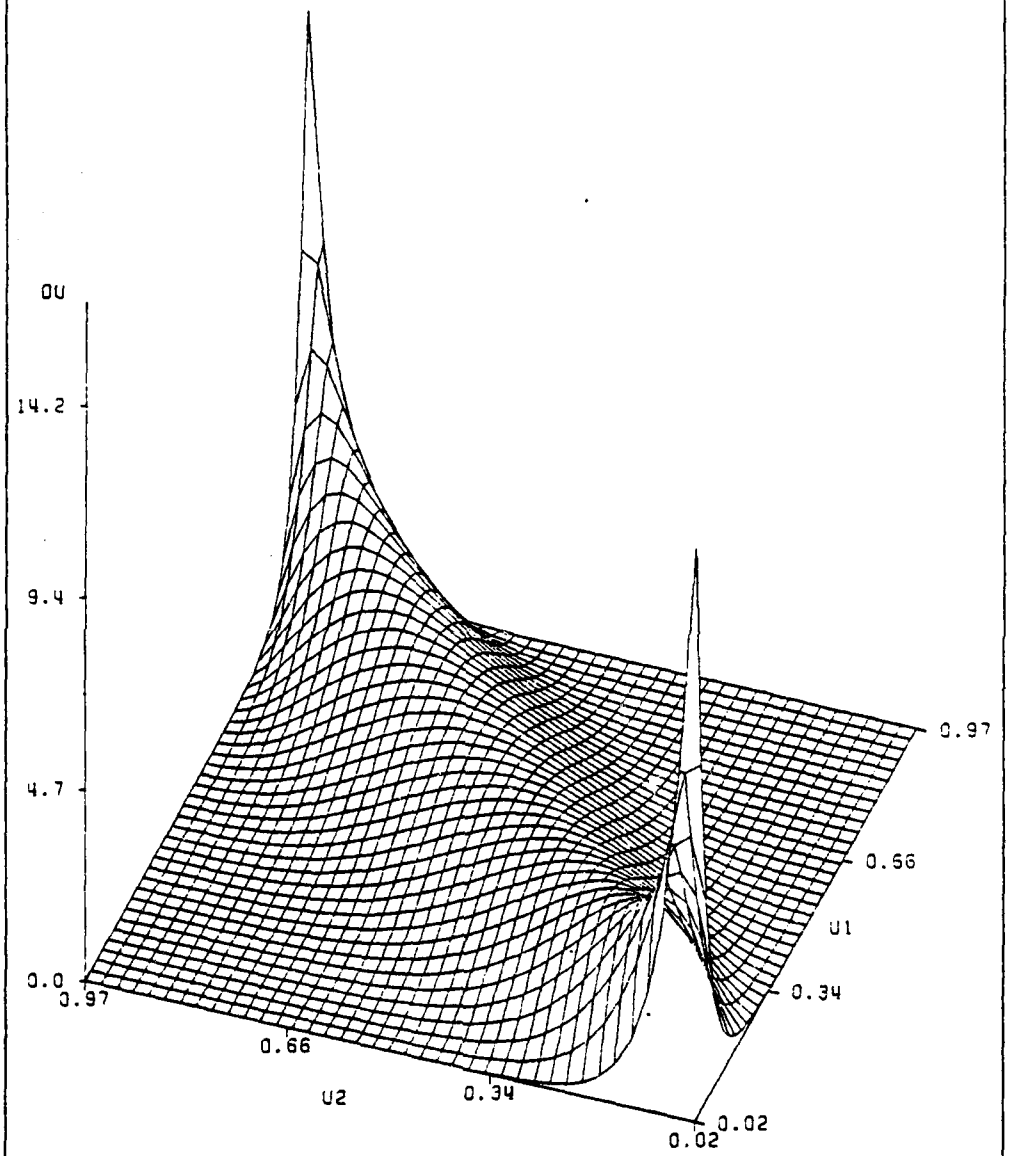


Figure 3. Bivariate Standard Normal Dependence Density for Rho=0.9

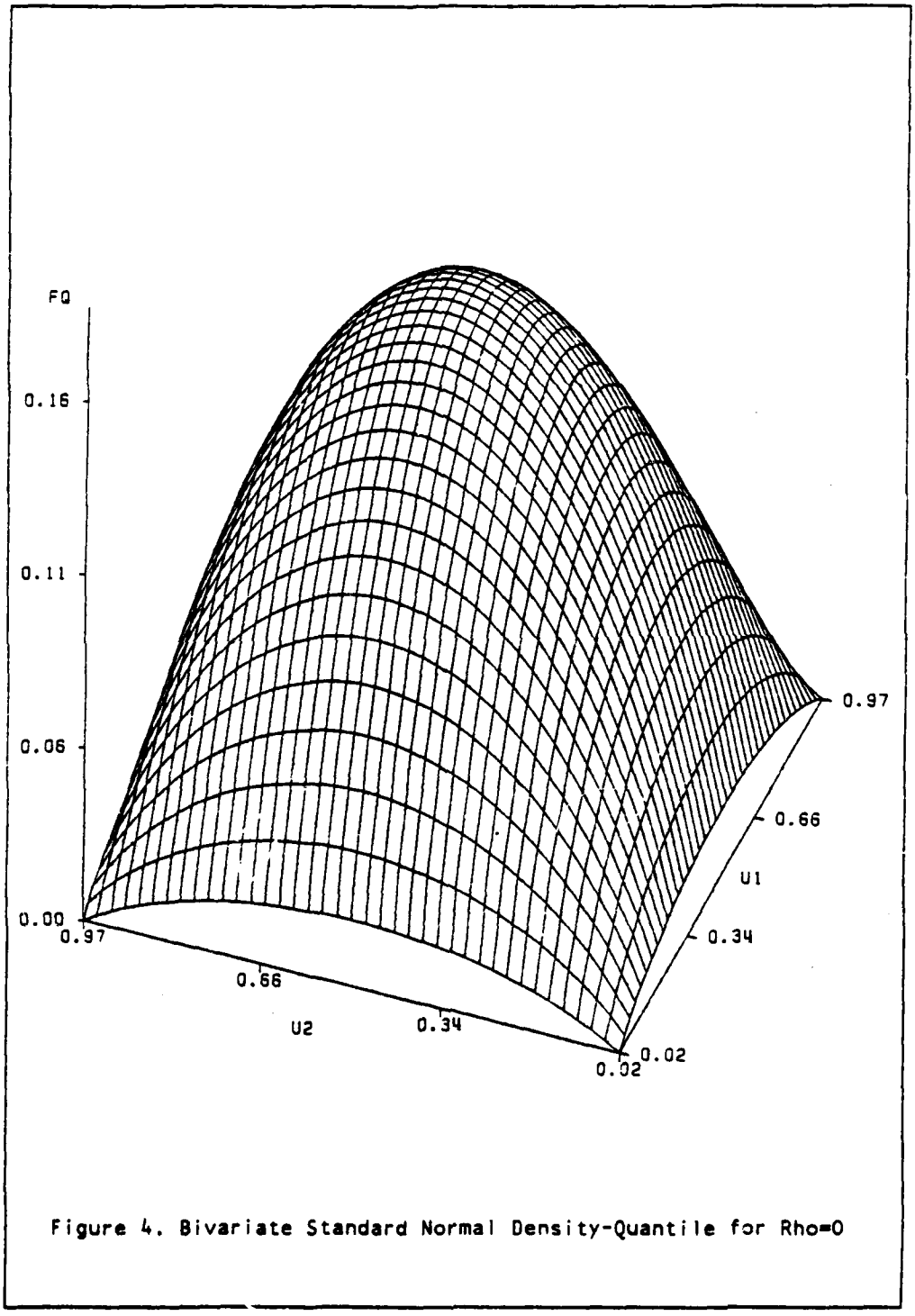


Figure 4. Bivariate Standard Normal Density-Quantile for Rho=0

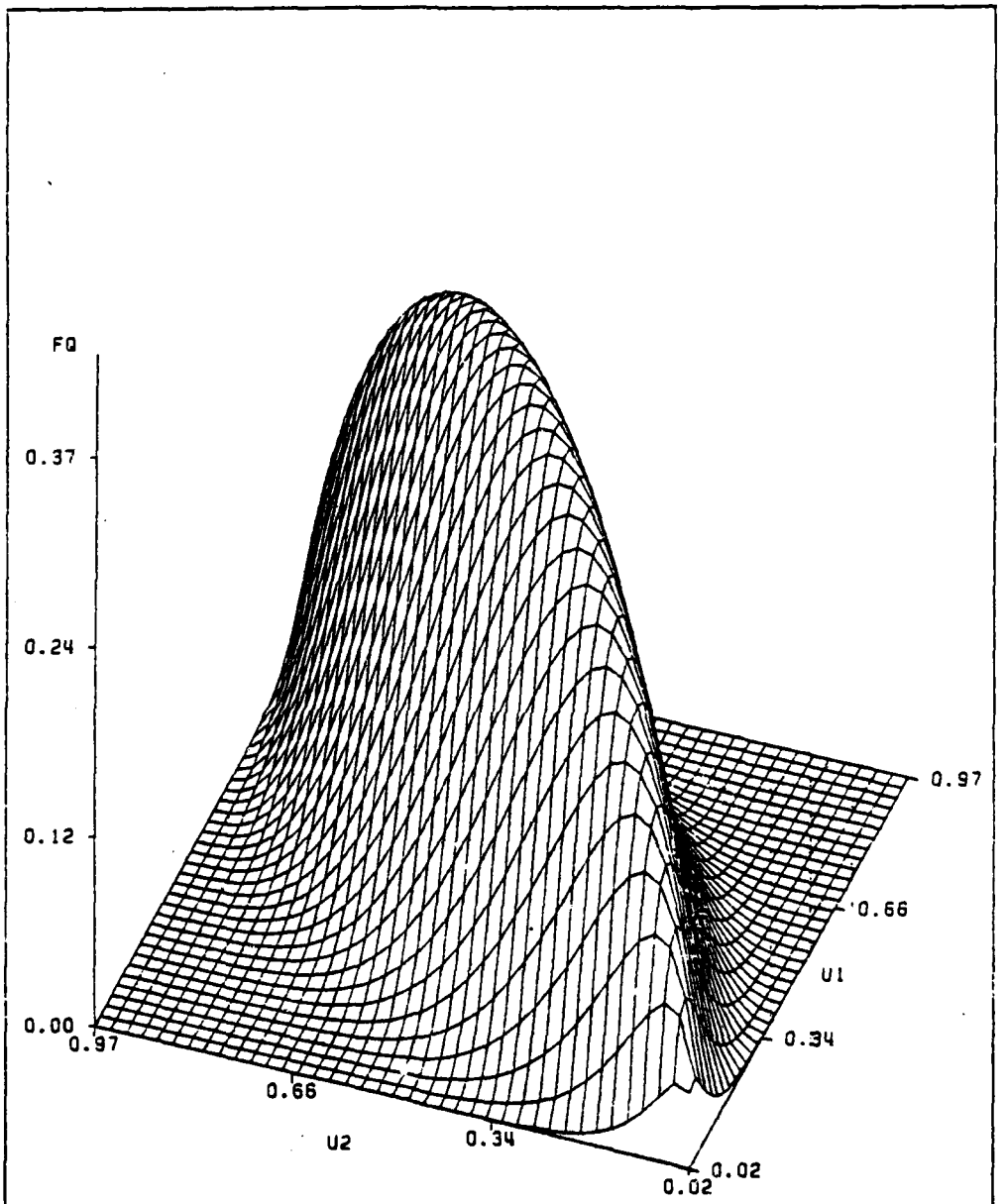


Figure 5. Bivariate Standard Normal Density-Quantile for Rho=0.9

Then

$$F_{X,Y} \sim G_{X,Y} \text{ iff } G_{X,Y}(x,y) = F_{X,Y}[F_X^{-1}G_X(x), F_Y^{-1}G_Y(y)]. \quad (2.2.18)$$

Example 2.2.1 Let $\Phi(\cdot)$ be the standard normal c.d.f. Then for ρ satisfying $0<\rho<1$,

$$F_{X,Y}(x,y) = \Phi(x)\Phi(y) + [1-\Phi(x)][1-\Phi(y)]$$

$$\{(\min[(1-\Phi(x)), (1-\Phi(y))])^{\rho} - 1\} \quad (2.2.19)$$

has standard normal marginals. This c.d.f. is derived using the bivariate uniform c.d.f

$$D(u_1,u_2) = u_1u_2 + (1-u_1)(1-u_2)$$

$$\{(\min[(1-u_1), (1-u_2)])^{\rho} - 1\} \quad (2.2.20)$$

and taking advantage of (2.2.18). Mardia (1970) discusses this c.d.f. in connection with the bivariate exponential distribution proposed by Marshall and Olkin (1967).

Kimeldorf and Sampson (1975a) discuss one-parameter families of bivariate distributions in light of Theorem 2.2.3. $G_{X,Y}$ of (2.2.18) *is called a* (G_X, G_Y)*-translate of* $F_{X,Y}$. *Ideally, one-parameter families of bivariate distributions exhibit a parameter that provides*

some measure of association between the random variables. Such is the case for the one-parameter family of bivariate standard normal distributions (i.e., $\mu_x = \mu_y = 0, \sigma_x = \sigma_y = 1$). Kimeldorf and Sampson base their discussion on uniform representations of bivariate distributions.

The uniform density $d(u_1, u_2)$ defined by (2.2.17) also has applications to regression problems. Using the definition of $E(Y|X)$, one may write the equivalent expression

$$E[Y|X=Q_X(u)] = \int_0^1 Q_Y(u_2)\, d(u_1, u_2)\, du_2 \qquad (2.2.21)$$

by using the change of variable $X=Q_X(U_1)$ and $Y=Q_Y(U_2)$ where U_1 and U_2 are (possibly dependent) uniform (0,1) random variables. Equation (2.2.21) justifies Parzen's use of the term regression density for $d(u_1, u_2)$. Furthermore, this representation suggests applications to nonparametric regression which will be considered in Chapter 4.

One may also consider general quantile representations of conditional probability functions. Corresponding to the conditional c.d.f. $F_{Y|X}(y|x)$ is a conditional quantile function $Q_{Y|X}(u|x)$ defined by (2.2.1). Parzen (1977) uses the usual change of variable to obtain

$$F_{Y|X}(y|x) = \int_0^{u_2} d(F_X(x), u_2')\, du_2' \qquad (2.2.22)$$

where $u_2 = F_Y(y)$. He then expresses $Q_{Y|X}(u|x)$ in terms of the unconditional quantile function $Q_X(\cdot)$ evaluated at an inverse representation of the right hand side of (2.2.22). One would prefer a

simpler expression conducive to estimation from sample data, but Parzen suggests estimators for $Q_{Y|X}$ whose properties remain to be investigated. The value of the conditional quantile function is illustrated in the following application.

To generate a univariate random sample of size n with specified c.d.f. F, one generates a uniform random sample $U_1,\ldots,U_n$ and forms

$$X_i = Q(U_i),\ i=1,\ldots,n, \tag{2.2.23}$$

where Q is the quantile function corresponding to F. Theorem 2.2.1 guarantees that the sample has common c.d.f. F. The extension of this approach to the bivariate case, however, is not obvious. The usual approach is to generate collections of random variables $V_1,V_2,\ldots,V_k$ and form

$$X=g(V_1,V_2,\ldots,V_k),\ Y=h(V_1,V_2,\ldots,V_k) \tag{2.2.24}$$

so that the appropriate distribution theory guarantees that X and Y have specified joint c.d.f. $F_{X,Y}$. This entails generating kn random variables to obtain 2n random variables. Furthermore, the V values usually are transformed uniform values based on (2.2.23) so that the simulation problem becomes unreasonably complicated. In some cases, the appropriate distribution theory does not exist to generate the desired random sample. To overcome this problem, one may develop a general procedure based on the conditional quantile function.

Let U_1 and U_2 be independent uniform (0,1) random variables.

Specify joint c.d.f. $F_{X,Y}$ and form

$$X=Q_X(U_1),\ Y=Q_{Y|X}(U_2|X) \tag{2.2.25}$$

where the quantile functions correspond to the choice of c.d.f. $F_{X,Y}$. Then X and Y have specified joint c.d.f. $F_{X,Y}$. To generate a sample of size n, one merely generates two independent samples of uniform random variables and uses (2.2.25) to form the corresponding bivariate sample with c.d.f. $F_{X,Y}$. Kennedy and Gentle (1980) consider a variety of techniques for generating uniform (0,1) random variables and discuss the design of Monte Carlo experiments. Such techniques will be applied in Chapter 4.

One attempts to estimate the above quantities with statistics that possess desirable properties. An important sample object useful in developing statistics of interest is the <u>empirical distribution function</u> (e.d.f.) $F_n(x)$ defined by

$$F_n(x) = (1/n)\{\text{no. of data points} \le x\}. \tag{2.2.26}$$

Formally, one assumes a collection $X_1,X_2,\ldots,X_n$ of i.i.d. r.v.'s, i.e., a random sample of size n, and defines the empirical c.d.f. $F_n(x)$ by

$$F_n(x) = (1/n)\sum_{i=1}^{n} I_A(X_i) \tag{2.2.27}$$

where $A=(-\infty,x]$ and

$$I_A(t) = \begin{cases} 1 \text{ if } t \in A, \\ 0 \text{ if } t \notin A, \end{cases} \qquad (2.2.28)$$

is the indicator function. One observes immediately that $F_n(x)$ satisfies the properties of a c.d.f., i.e., nondecreasing, continuous from the right, $0 \le F_n(x) \le 1$, $F_n(-\infty)=0$, and $F_n(\infty)=1$. Furthermore, $F_n(x)$ may be thought of as a stochastic process although presently attention is paid to $F_n(x)$ for fixed x.

We now briefly state some important properties of $F_n(x)$. Let $F(x)$ be the true population c.d.f. generating the data. Then

$$E[F_n(x)] = F(x), \quad \mathrm{Var}[F_n(x)] = F(x)[1-F(x)]/n. \qquad (2.2.29)$$

Thus, $F_n(x) \to F(x)$ as $n \to \infty$ in quadratic mean, or $F_n(x)$ is consistent in mean square for estimating $F(x)$.

The representation (2.2.27) also permits direct application of the Strong Law of Large Numbers (SLLN) and the Central Limit Theorem (CLT). One notes that $nF_n(x)$ is exactly binomially distributed with parameters $(n, F(x))$, which makes it easy to deduce that $F_n(x)$ is strongly consistent for estimating $F(x)$ and that $F_n(x)$ suitably standardized is asymptotically normally distributed. Note that these results pertain to the pointwise estimation of $F(x)$ by $F_n(x)$. Global measures characterizing the "closeness" of F_n in approximating F may be found in Durbin (1973) with important asymptotic results stated therein.

The Lebesgue-Stieltjes integral w.r.t. the empirical c.d.f. is

often employed to obtain method of moments estimates for the corresponding population parameter. For example, for the parameter defined by

$$\mu = \int_{-\infty}^{\infty} x \, dF(x), \tag{2.2.30}$$

the corresponding method of moments estimator $\bar{X}_n$ may be represented by

$$\bar{X}_n = \int_{-\infty}^{\infty} x \, dF_n(x). \tag{2.2.31}$$

This approach to obtaining estimators has many applications. Some of the information parameters of the next section will be estimated using this approach.

An empirical function fundamental to the quantile approach is the <u>empirical quantile function</u> given by

$$Q_n(u) = F_n^{-1}(u) = \inf\{x: F_n(x) \geq u\} \tag{2.2.32}$$

Upon closer inspection one realizes that (2.2.32) is equivalent to

$$Q_n(u) = X_{(j)} \text{ for } (j-1)/n < u \leq j/n, \; j=1,\ldots,n, \tag{2.2.33}$$

where $X_{(j)}$ is the j-th <u>order statistic</u> of the random sample. Parzen (1979b) suggests that $Q_n(0)$ be taken to be a natural minimum when one is available. If $Q(u)$ is the true population quantile function, one may define the <u>sample quantile process</u> by

$$A_n(u) = \sqrt{n}[Q_n(u) - Q(u)], \quad 0 \le u \le 1. \tag{2.2.34}$$

The following results may be found in Csörgö and Révész (1981).

Theorem 2.2.4 Let u, $0<u\le 1$, be given. Let $F(x)$ be absolutely continuous in an interval about $Q(u)$, and let $fQ(u)$ be positive and continuous at u. Then as $n\to\infty$,

$$fQ(u)\, A_n(u) / \sqrt{u(1-u)} \xrightarrow{d} N(0,1). \tag{2.2.35}$$

Theorem 2.2.5 Let $Q(u)$ be continuous at u. Then as $n\to\infty$,

$$Q_n(u) \xrightarrow{a.s.} Q(u). \tag{2.2.36}$$

Further results are given in Csörgö and Révész (1981). Serfling (1980) also provides similar results for the sample quantile function. More general results treat $A_n(u)$ as a stochastic process and exhibit the left hand side of (2.2.35) as converging in distribution to a Brownian Bridge stochastic process. The definition of a Brownian Bridge is given in Section 2.5, but for now, one notes that the results of Theorem 2.2.4 may be generalized to conclude

$$fQ(u)\, A_n(u) \xrightarrow{d} B(u), \text{ for all } u, \tag{2.2.37}$$

where B(u) is a Brownian Bridge process.

One may prefer to use the piecewise linear definition of the sample quantile function given by

$$Q_n(u) = n[(j/n)-u]X_{(j-1)} + n[u-(j-1)/n]X_{(j)},$$

$$(j-1)/n \le u \le j/n, j=1,\ldots,n, \quad (2.2.38)$$

or the shifted piecewise linear version

$$Q_n(u) = n[(j+.5)/n-u]X_{(j)} + n[u-(j-.5)/n]X_{(j+1)},$$

$$(j-.5)/n \le u \le (j+.5)/n, j=1,\ldots,n. \quad (2.2.39)$$

Using definition (2.2.39) suggests that the empirical quantile-density be defined by

$$q_n(u) = n(X_{(j+1)}-X_{(j)}), \quad (j-.5)/n<u<(j+.5)/n, \quad j=1,\ldots,n-1. \quad (2.2.40)$$

This definition of $q_n(u)$ is the derivative of (2.2.39). For any definition of $Q_n(u)$, one may take the corresponding $q_n(u)$ to be the raw derivative

$$q_n(u) = [Q_n(u+h)-Q_n(u-h)]/(2h), \quad 0<u<1, \quad (2.2.41)$$

where $h=h(n)$ is some predetermined positive function of n. Bloch and

Gastwirth (1968) use (2.2.41) corresponding to $Q(u)$ defined by (2.2.33). Vasicek (1976) applies this definition to obtain the g.o.f. test of normality discussed previously. Observe that for any well defined $q_n(u)$, $fQ_n(u)=1/q_n(u)$ is an estimate of the density-quantile function by virtue of the reciprocal identity. The g.o.f. statistic of Vasicek uses this fact to define the sample entropy discussed in the next section.

A problem with the $q_n(u)$ estimates above is that they are not necessarily consistent estimators of $q(u)$. Hence, one usually seeks "smoothed" or corrected versions that yield nice asymptotic results. One notes that often $q(u)$ is not a function of interest except as it is applied using the reciprocal identity. Thus, techniques for estimating $q(u)$ are employed and if estimates of $fQ(u)$ are desired, one then applies the reciprocal identity. The estimation of $fQ(u)$ will be considered in Chapter 3.

The bivariate functions of interest are $F_{X,Y}(x,y)$, $f_{X,Y}(x,y)$, $D(u_1,u_2)$, $d(u_1,u_2)$, and $Q_{Y|X}(y|x)$. Raw estimates of $F_{X,Y}(x,y)$ and $D(u_1,u_2)$ may be obtained analogously to the empirical c.d.f., i.e., defined with jumps of size $1/n$ at the points (X_i,Y_i) and $(Q_i/(n+1),R_i/(n+1))$ respectively, where $Q_i=\text{rank}(X_i)$ and $R_i=\text{rank}(Y_i)$. Improved versions of these estimators will be considered in Chapters 3 and 4. Parzen (1977) suggests techniques for estimating $Q_{Y|X}(y|x)$ based on raw estimates of $D(u_1,u_2)$. This subject will be discussed further in Chapter 4 in relation to several techniques of bivariate density estimation. The asymptotic results for the bivariate case, however, remain to be investigated.

2.3 Information and Entropy

The concept of statistical information or information numbers has many useful applications in statistical analysis (Kullback, 1978, Rao, 1973). Fisher's information has been studied extensively and is of primary importance in uniform minimum variance estimation and maximum likelihood estimation. We will consider an alternative measure of information proposed by Shannon (1948) and studied in a statistical setting by Kullback (1978). The following definitions pertain to a measure of information and related concepts.

Definition 2.3.1 The information $I(f;g)$ between two densities $f(x)$ and $g(x)$ is given by

$$I(f;g) = \int_{-\infty}^{\infty} \{\log[f(x)/g(x)]\} f(x)\, dx. \qquad (2.3.1)$$

Definition 2.3.2 The entropy of a density $f(x)$ is given by

$$H(f) = \int_{-\infty}^{\infty} \{-\log f(x)\} f(x)\, dx. \qquad (2.3.2)$$

Definition 2.3.3 The cross-entropy between two densities $f(x)$ and $g(x)$ is given by

$$H(f;g) = \int_{-\infty}^{\infty} \{-\log g(x)\} f(x)\, dx. \qquad (2.3.3)$$

One immediately notes that $H(f)=H(f;f)$ and that

$$I(f;g) = H(f;g) - H(f). \tag{2.3.4}$$

Kullback (1978) proves the following fundamental theorem.

Theorem 2.3.1 Let $f(x)$ and $g(x)$ be probability densities. Then

$$I(f;g) \geq 0 \text{ a.e.} \tag{2.3.5}$$

Equation (2.3.5) is called the information inequality. We will exploit this inequality in constructing tests for independence between two random variables.

Generally, information is considered as a "distance" between two densities, although it is not a metric since it does not satisfy the triangle inequality. If one wishes a symmetric measure of information, one such definition is provided by

$$J(f;g) = I(f;g) + I(g;f). \tag{2.3.6}$$

Observe,

$$J(f;g) = \int_{-\infty}^{\infty} [f(x)-g(x)] \log[f(x)/g(x)]\, dx. \tag{2.3.7}$$

Kullback calls $J(f;g)$ the divergence between f and g.

One may also note that

$$I(f;g) = E_f[\log f(X)] - E_f[\log g(X)] \tag{2.3.8}$$

and that

$$H(f) = E_f[-\log f(X)]. \tag{2.3.9}$$

These expectations need not be finite.

More general measures of information may also be developed. Parzen (1982) discusses several general information measures, in particular the bi-information given by

$$II(f;g) = \int_{-\infty}^{\infty} |\log[f(x)/g(x)]|^2 f(x)\, dx. \tag{2.3.10}$$

We will exploit this definition as an estimation criterion in Chapter 4.

Our main application of information as a statistical measure will be to the problem of ascertaining whether two random variables X and Y are associated. For joint p.d.f. $f_{X,Y}$ and marginals f_X, f_Y, one obtains (see section 4.5)

$$I(f_{X,Y};f_X f_Y) = -H(d) \tag{2.3.11}$$

where d is the dependence density for X and Y. One may then exploit (2.3.11) as a measure of dependence or association. Linfoot (1957) was one of the first authors to consider information as a measure of association between X and Y. Of primary importance, however, is the fact that X and Y are independent if and only if $I(f_{X,Y};f_X f_Y)=0$ by virtue of the information inequality of Theorem 2.3.1.

One may desire to emphasize the modeling of probability laws of random variables by using the alternate notation

$$I(Y|X) = I(f_Y; f_{Y|X}). \tag{2.3.12}$$

Using such notation it is easy to show that

$$I(Y|X) = H(Y) - H(Y|X) \tag{2.3.13}$$

where

$$H(Y|X) = H(f_{Y|X}). \tag{2.3.14}$$

These results are readily applicable to regression and prediction problems.

Information thus has a dual role in statistics being a parameter of interest or a criterion function depending upon the setting for the problem of interest. We will apply information criterion functions to the problem of density estimation in section 4.4. The use of information as a dependence parameter will be investigated in section 4.5. Generally, information serves as a useful goodness-of-fit criterion also. Consider the general Neyman-Pearson theory of hypothesis testing. Recall, one rejects H_o: $f(x)=f_o(x)$ in favor of H_1: $f(x)=f_1(x)$ for specified α if

$$\prod_{i=1}^{n} f_1(X_i) \geq k \prod_{i=1}^{n} f_o(X_i), \tag{2.3.15}$$

where k is chosen to satisfy

$$P\{[\prod_{i=1}^{n} f_1(X_i)]/[\prod_{i=1}^{n} f_o(X_i)] \geq k | H \} = \alpha . \qquad (2.3.16)$$

Taking logarithms of (2.3.15) and simplifying, one obtains the equivalent expression

$$(1/n) \sum_{i=1}^{n} \log f_1(X_i) - (1/n) \sum_{i=1}^{n} \log f_o(X_i) \geq k , \qquad (2.3.17)$$

which can be written

$$H_{F_n}(f_1) - H_{F_n}(f_o) \geq k , \qquad (2.3.18)$$

where

$$H_{F_n}(f) = \int_{-\infty}^{\infty} -\log f(x) \, dF_n(x) . \qquad (2.3.19)$$

Another expression equivalent to (2.3.15) is

$$I_{F_n}(f_1;f_o) \geq k \qquad (2.3.20)$$

where

$$I_{F_n}(f;g) = \int_{-\infty}^{\infty} \log[f(x)/g(x)] \, dF_n(x) . \qquad (2.3.21)$$

Vasicek (1976) develops an entropy based test of normality with

critical region defined by

$$H_n(\hat{fQ}_n) \leq H_\alpha(m,n) \qquad (2.3.22)$$

where $H_n(\hat{fQ}_n)$ given by

$$H(\hat{fQ}_n) = (1/n) \sum_{i=1}^{n} \log\{(n/2m)(X_{(i+m)} - X_{(i-m)})\}, \qquad (2.3.23)$$

with $X_{(i)} = X_{(1)}$ for $i<1$ and $X_{(i)} = X_{(n)}$ for $i>n$, is the sample entropy of a nearest neighbor estimate of the density-quantile function and $H_\alpha(m,n)$ is the corresponding critical value for significance level α. Dudewicz and Van Der Meulen (1981) investigate power properties of this procedure and conclude in simulation studies that the test is compet tive with existing g.o.f. procedures. Vasicek shows that the sample entropy is consistent for estimating H(f). A similar procedure for the bivariate case will be considered in Chapter 4.

Since the bivariate normal is of special interest, we note that the information between the joint p.d.f. and the product of the marginals for this special case is given by

$$I(f_{X,Y}; f_X f_Y) = -.5 \log(1-\rho^2) \qquad (2.3.24)$$

so that the information in this case is a function only of the correlation coefficient ρ.

The parametric approach to statistical inference using information theory, with emphasis on classical normal theory, has been

studied extensively in the literature with Kullback (1978) providing a fundamental reference. Only recently has information theory been applied to nonparametric problems and exploratory data analysis. This work attempts to contribute to the application of information theory to such statistical problems.

2.4 Some Fundamental Concepts from Approximation Theory

Approximation theory has as its primary goal the approximation of a function (or a graph or a curve). Several examples may illuminate the need for such an approximation.

Example 2.4.1 The error function defined by

$$\text{erf}(x) = (2/\sqrt{\pi}) \int_0^x \exp(-y^2)\, dy \qquad (2.4.1)$$

cannot be obtained explicitly for specified x since the integral on the right hand side of (2.4.1) cannot be simplified. Hence, one seeks to approximate erf(x) by approximating the integral for given x. One solution is to employ numerical integration techniques to approximate erf(x). Statisticians are interested in this problem because the standard normal c.d.f. $\Phi(x)$ may be expressed by

$$\Phi(x) = .5 + .5\ \text{erf}(x/\sqrt{2}),\ x>0. \qquad (2.4.2)$$

Example 2.4.2 Let $y(x,t)$ represent the displacement of a tightly stretched string at time t vibrating in the xy-plane. The analysis of the problem of the vibrating string yields the differential equation

$$\frac{\partial^2 y(x,t)}{\partial t^2} = k^2 \frac{\partial^2 y(x,t)}{\partial x^2} \qquad (2.7.3)$$

also called the wave equation. A simple (closed form, finite) expression for $y(x,t)$ has not been obtained for specific boundary conditions. One solution is to express $y(x,t)$ as an infinite Fourier series and truncate at a suitable order (see, e.g., Churchill, 1969).

These two examples depict a situation where an exact numerical value cannot be obtained for a function at a specified point because no "simple" expression for the function has been discovered. Abramowitz and Stegun (1972) provide a useful reference for such problems, especially when one seeks to obtain approximations with specified bounds on precision.

The following two examples depict settings in which approximation theory and statistical estimation theory seem to become entwined.

Example 2.4.3 Given a set of bivariate data $(X_1,Y_1),\ldots,(X_n,Y_n)$, the relationship

$$Y_i = r(X_i,\underline{\theta}) + \varepsilon i \qquad (2.4.4)$$

is known to hold where $r(\cdot,\cdot)$ is a specified function of X and

parameters $\underline{\theta} = (\theta_1, \ldots, \theta_k)$ and $\{\varepsilon_i\}$ are i.i.d. r.v.'s with known distribution. The function $r(x, \underline{\theta})$ is known except for the parameters. One seeks to approximate r by estimating the parameters based on the sample data.

Example 2.4.4 A set of data is generated by a probability mechanism with probability density function f. If f is unknown, one seeks to approximate f based on the observations in the sample data. Chapter 3 presents several solutions to this problem.

Example 2.4.3 illustrates how approximation theory and statistical estimation theory compliment each other. However, Example 2.4.4 seems to be an exercise in statistical estimation alone. One may find it difficult to distinguish between the terms "approximation" and "estimation". In this work, approximation theory will refer to the concepts and theorems employed to approximate a function with known mathematical properties. The approximation may be obtained by rigorous mathematical arguments or by working with criterion functions and sample data. Estimation theory, on the other hand, must always resort to sample data and hence has a stochastic element not essential to approximation theory. In the context of Example 2.4.4, estimation theory would treat f(x) as a "parameter" while approximation theory would treat it as a function with specified properties. Any ambiguity in the use of these terms should pose no serious obstacles in applying them to problems of interest.

To approximate a function one usually must restrict f to belong

to a certain class of functions. Theory is then developed to approximate functions in a given class. More generally, one may consider a space of functions whose members possess certain properties. The simplest class of functions might be the space of constant functions, while one of the more complex classes might consist of measurable functions. To begin the study of approximating elements in a space of functions, several concepts will be introduced that will be useful in defining function spaces.

Definition 2.4.1 A metric space (M,d) is a nonempty set M of elements together with a real-valued function $d:M\times M\to R$ called a metric that satisfies the following properties for all $x,y,z\in M$:

i) $d(x,y)\geq 0$;

ii) $d(x,y)=0$ iff $x=y$;

iii) $d(x,y)=d(y,x)$, and

iv) $d(x,y)\leq d(x,z)+d(z,y)$.

The function d is also called a distance, and property (iv) above is called the triangle inequality because of its application to Euclidean 2-space with the metric $d(x,y)=|x-y|$.

Definition 2.4.2 A vector space (or linear space) over the reals is a set of elements (called vectors) V together with two operations (functions) $+:V\times V\to V$ and $\cdot:R\times V\to V$ which satisfy the following properties for all $x,y,z\in V$ and $\lambda,\mu\in R$:

i) $x+y=y+x$;

ii) $(x+y)+z=x+(y+z)$;

iii) there exists $\theta \in V$ such that $x+\theta=x$ for all $x \in V$;

iv) $\lambda(x+y)=x+\lambda y$;

v) $(\lambda+\mu)x=\lambda x+\mu x$;

vi) $\lambda(\mu x)=(\lambda\mu)x$; and

vii) $0 \cdot x=$, $1 \cdot x=x$.

The set of real numbers in this setting is also called a set of scalars, hence in general one speaks of a vector space and a set of scalars.

Definition 2.4.3 A real-valued function $\|\cdot\|: V \to R$ defined on a vector space V is called a norm if the following properties are satisfied for all $x,y \in V$ and $\lambda \in R$:

i) $\|x\| \geq 0$;

ii) $\|x\|=0$ iff $x=\theta$;

iii) $\|x+y\| \leq \|x\|+\|y\|$; and

iv) $\|\lambda x\|=|\lambda|\|x\|$.

A vector space that possesses such a norm is called appropriately enough a normed vector space.

One immediately notes that $d(x,y)=\|x-y\|$ defines a metric and hence a normed vector space is also a metric space. The concept of a

distance measure or metric gives one a firm grasp on many abstract concepts.

Definition 2.4.4 Let $\{x_n\}$ be a sequence of vectors in a normed vector space V with norm $\|\cdot\|$. The sequence $\{x_n\}$ is called a Cauchy sequence if for given $\varepsilon>0$, there exists an N such that for all $n\geq N$ and $m\geq N$, $\|x_n - x_m\| < \varepsilon$.

Definition 2.4.5 A normed vector space V is said to be complete if all Cauchy sequences converge, i.e., if $\{x_n\}$ is a Cauchy sequence, then there exists $x\in V$ such that $\lim_{n\to\infty} x_n = x$. A complete normed vector space is called a Banach space.

Definition 2.4.6 A set H is called an inner product space if it is a vector space and if there exists a real-valued function $(\ ,\ ): H\times H \to R$ called an inner product that satisfies the following properties for all $x,y,z \in H$ and all $\lambda\in R$:

i) $(\lambda x, y) = \lambda(x,y)$;

ii) $(x+y,z) = (x,z)+(y,z)$;

iii) $(x,y) = (y,x)$; and

iv) $(x,x) > 0$ if $x\neq\theta$.

One may permit the inner product to be complex valued, i.e., $(\cdot,\cdot): H\times H \to C$, in which case property (iii) above becomes

iii) $(x,y)=\overline{(y,x)}$

where $\bar{z}$ is the complex conjugate of z. Note that $\|x\|^2=(x,x)$ defines a norm so that an inner product space is also a normed vector space.

Definition 2.4.7 A complete inner product space is called a Hilbert space. (Some authors, e.g., Davis, 1975, give a more restrictive definition of a Hilbert space, but this definition seems fairly standard in the literature. See, e.g., Loève, 1977, Royden, 1968, or Hewitt and Stromberg, 1965.)

When one moves down the hierarchy of spaces defined above, the transition from inner product to norm to metric will be understood to follow the convention given unless otherwise noted. The hierarchy is emphasized as follows:

Hilbert space → Banach space → normed vector space → metric space.

While restricting functions to belong to a Hilbert space may seem severe, one will soon discover that many functions of statistical importance fit nicely into a special class of Hilbert spaces called the L^p spaces.

Theorem 2.4.1 Let H be a Hilbert space with inner product $(\cdot,\cdot)$. Then for any $x,y\in H$,

$$(x,y) \leq \|x\| \, \|y\|. \tag{2.4.5}$$

The inequality (2.4.5) is known as the Cauchy-Schwarz inequality.

Example 2.4.5 For two random variables X and Y possessing finite second moments, $Cov(X,Y)$ is an inner product. Thus, $\sqrt{Var(X)} = \sqrt{Cov(X,X)}$ is a norm, and hence

$$Cov(X,Y) \leq \sqrt{Var(X)\ Var(Y)}. \tag{2.4.6}$$

This is the common statistical version of the Cauchy-Schwarz inequality.

Definition 2.4.8 Let H be a Hilbert space with inner product $(\cdot,\cdot)$ and let $x,y \in H$. One says that x and y are orthogonal if $(x,y)=0$. A set $S \in H$ is called an orthogonal system if for all distinct elements $x,y \in S$, $(x,y)=0$. Furthermore, if $\|x\|=1$ for all $x \in S$, S is called an orthonormal system.

Remark 2.4.1 If H is separable (see, e.g., Royden, 1968), then every orthonormal system in H must be countable. This work will need only consider separable Hilbert spaces. Theorem 2.4.3 below will illustrate why.

Definition 2.4.9 Let H be a separable Hilbert space and let $\{\phi_k\}_{k=1}^{\infty}$ be an orthonormal system in H. The Fourier coefficients

w.r.t. $\{\phi_k\}$ of an element $x \in H$ are defined by

$$\theta_k = (x, \phi_k) . \tag{2.4.7}$$

Theorem 2.4.2 Let H be a separable Hilbert space and $\{\phi_k\}_{k=1}^{\infty}$ be the Fourier coefficients w.r.t. $\{\phi_k\}_{k=1}^{\infty}$ of an element $x \in H$. Then

$$\sum_{k=1}^{\infty} \theta^2 \leq \|x\|^2 . \tag{2.4.8}$$

This is known as Bessel's inequality.

Deinition 2.4.10 Let H be a separable Hilbert space and let $\{\phi_k\}_{k=1}^{\infty}$ be an orthonormal system in H. If $(x, \phi_k)=0$ for all k implies that $x=\theta$ (here θ is the identity element), then $\{\phi_k\}_{k=1}^{\infty}$ is said to be a complete orthonormal system.

The justification for the term "complete" becomes evident in the following theorem. Royden (1968, p. 212) also gives motivation for this usage.

Theorem 2.4.3 Let H be a separable Hilbert space. Then every orthonormal system in H is countable and there exists a complete orthonormal system. If $\{\phi_k\}_{k=1}^{\infty}$ is any complete orthonormal system in H and $x \in H$, then

$$x = \sum_{k=1}^{\infty} \theta_k \phi_k \tag{2.4.9}$$

where $\theta_k = (x, \phi_k)$. Equation (2.4.9) is said to be the Fourier series representation of x and specifically means $\lim_{m \to \infty} \left\| x - \sum_{k=1}^{m} \theta_k \phi_k \right\| = 0$.

Remark 2.4.2 Every complete orthonormal system in a separable Hilbert space has the same number of elements. This number is called the dimension of H.

Remark 2.4.3 If the dimension of H is finite, then H is a finite dimensional vector space and any complete orthonormal system in H is a basis for H. One may also consider countably infinite basis sets if desired, in which case any complete orthonormal system in H is a basis for H. The Gramm-Schmidt orthonormalization process (see, e.g., Hewitt and Stromberg, 1965, pp. 240-242) permits construction of an orthonormal system given any basis set for a vector space.

Remark 2.4.4 In the context of Theorem 2.4.3, it follows that for $x \in H$,

$$\left\| x \right\|^2 = \sum_{k=1}^{\infty} \theta_k^2. \tag{2.4.10}$$

This is known as Parseval's identity.

Theorem 2.4.3 provides the foundation for many useful expansion techniques used to approximate a function of interest. One need only

decide upon an appropriate orthonormal system in a Hilbert space of functions to construct an approximation based upon equation (2.4.9). First, however, an appropriate Hilbert space must be identified that contains functions of interest.

Definition 2.4.11 A (Lebesgue) measurable function f is said to belong to the space $L^p(a,b)$ if

$$\| f \|_p = \left\{ \int_a^b |f(x)|^p dx \right\}^{1/p} < \infty \qquad (2.4.11)$$

for $1 \leq p < \infty$.

If one defines an inner product between two functions f and g by

$$(f,g) = \int_a^b f(x)\, \overline{g(x)}\, dx, \qquad (2.4.12)$$

where $\overline{g(x)}$ is the complex conjugate of $g(x)$, then the corresponding norm is given by

$$\| f \| = \left\{ \int_a^b |f(x)|^2 dx \right\}^{\frac{1}{2}} . \qquad (2.4.13)$$

It can be shown that $L^2(a,b)$ is a separable Hilbert space, and hence Theorem 2.4.3 applies for functions in $L^2(a,b)$. Let $\{\phi_k(x)\}_{k=1}^{\infty}$ be a complete orthonormal system in $L^2(a,b)$. Then for $f \in L^2(a,b)$, equation (2.4.9) becomes

$$f(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x) \tag{2.4.14}$$

where

$$\theta_k = (f,\phi_k) = \int_a^b f(x)\ \overline{\phi_k(x)}\ dx \tag{2.4.15}$$

are the Fourier coefficients w.r.t. $\{\phi_k(x)\}$.

Now, suppose one wishes to approximate a function $f \in L^2(a,b)$ by a suitable finite expression. One solution is to chose the truncated Fourier series representation

$$f_m(x) = \sum_{k=1}^{m} \theta_k \phi_k(x). \tag{2.4.16}$$

Indeed, this approximation for f has some nice properties.

Theorem 2.4.4 Let H be a separable Hilbert space and let $\{\phi_k\}_{k=1}^{\infty}$ be a complete orthonormal system in H. Then for any $x \in H$,

$$\| x - \sum_{k=1}^{m} (x,\phi_k)\phi_k \| \leq \| x - \sum_{k=1}^{m} \theta_k \phi_k \| \tag{2.4.17}$$

for any choice of constants $\theta_1,\ldots,\theta_m$.

Observe that Theorem 2.4.4 implies that the best approximator w.r.t. the least squares criterion for an element in a separable Hilbert space is provided by the truncated Fourier series representation.

If one seeks a geometric interpretation of least squares

approximation, some fundamental definitions and notation are required. Since a separable Hilbert space is also a vector space, one may employ an analogy to vectors in Euclidean space and define for any two elements x,y in a Hilbert space H the <u>projection</u> of x on y by

$$\text{proj}(x,y) = [(x,y)/(y,y)]y. \qquad (2.4.18)$$

Using this interpretation, one sees that an element of a separable Hilbert space may be expressed as the sum of the projections of the element on the elements of an orthonormal system. Furthermore, one observes that the orthonormal system $\{\phi_k\}_{k=1}^{\infty}$ defines a subspace of the separable Hilbert space H. In this sense, the truncated Fourier series representation for an element x in H is essentially the projection of x into an m-dimensional subspace of H. For a clearer exposition on the geometric interpretation of least squares estimation, see Chapter 8 of Davis(1975).

Many of the results stated above also hold if $\{\phi_k\}_{k=1}^{\infty}$ is merely a system of orthogonal elements, except that terms involving $\|\phi_k\|$ may have been omitted. (Recall, in the setting considered $\|\phi_k\|=1$.) These results are used extensively in the statistical literature, most notably in the study of linear models. However, in many statistical settings, finite-dimensional vector space theory is sufficient to handle most problems of interest.

Many fundamental analysis texts discuss orthonormal systems for the space $L^2(a,b)$ as well as for other spaces of functions. The most popular systems include Jacobi polynomials, trigonometric systems, and

complex exponentials. Lanczos (1956), Rainville (1960), Davis (1975), and Powell (1981) are some basic references on approximation theory that identify various useful orthonormal systems.

For a discussion of some basic integration theorems and other results for L^P spaces, one may consult basic texts such as Royden (1968) or Hewitt and Stromberg (1965).

Most of the discussion thus far has emphasized only approximations by orthogonal expansion. Some results found in Bochner (1955) are valuable for other types of approximation. Parzen (1962) contains some useful essentials of approximation theory relevant to kernel density estimation based on some of the elements of approximation theory discussed by Bochner.

Theorem 2.4.5 Let $K(x)$ be a Borel measurable function satisfying

i) $\sup_x |K(x)| < \infty$;

ii) $\int_{-\infty}^{\infty} |K(x)|\, dx < \infty$; and

iii) $\lim_{x \to \infty} |x\, K(x)| = 0$.

Let $f(x)$ satisfy

iv) $\int_{-\infty}^{\infty} |f(x)|\, dx < \infty$.

Let $\{h(n)\}$ be a sequence of positive constants satisfying

v) $h(n) \to 0$ as $n \to \infty$.

Define approximating functions $f_n(x)$ by

$$f_n(x) = \int_{-\infty}^{\infty} [1/h(n)] K[y/h(n)] f(x-y)\ dy. \qquad (2.4.19)$$

Then at every continuity point x of f,

$$f_n(x) \xrightarrow{q.m.} f(x) \int_{-\infty}^{\infty} K(y)\ dy \quad \text{as } n \to \infty. \qquad (2.4.20)$$

Result (2.4.20) illustrates why one usually makes the additional restriction that K(x) integrates to one. One calls K a <u>kernel function</u>, and specific suggestions for K may be found in Bochner (1955) or Parzen (1962).

Parzen considers analogous results for the Fourier transforms

$$k(u) = \int_{-\infty}^{\infty} \exp(-iux) K(x)\ dx \qquad (2.4.21)$$

and

$$\phi(u) = \int_{-\infty}^{\infty} \exp(iux) f(x)\ dx, \qquad (2.4.22)$$

and extends these results to the solution of problems of density estimation. Some of these results will be mentioned in section 3.3.

Other results in approximation theory are applicable to statistics. As suggested, a reference such as Abramowitz and Stegun

(1972) is very handy for computer implementation of approximation theory results, and most such references require little mathematical expertise. We have avoided discussion of spline approximation techniques as we will continue to do throughout this work, but useful references such as Ahlberg, Nilson, and Walsh (1967) and Wahba (1971) adequately discuss the topic.

2.5 Some Fundamental Stochastic Processes and Complex Regression

The standard linear regression model is usually written

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ki} + \varepsilon_i \ , \ i=1,\ldots,n, \qquad (2.5.1)$$

where observations on the vector $\underline{X}=(X_1,\ldots,X_p)$ are assumed to be measured without error and the ε_i are uncorrelated random variables with common mean zero and common positive finite variance σ^2. In matrix notation, one writes

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \ , \qquad (2.5.2)$$

or one may express (2.5.1) by

$$E(Y|X=x) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k \ ,$$

$$\mathrm{Var}(Y|X=x) = \sigma^2. \qquad (2.5.3)$$

The well known Gauss-Markov Theorem (see, e.g., Graybill, 1976) states that under these equivalent model specifications, the least squares estimate of $\underline{\beta}$, namely

$$\hat{\underline{\beta}} = (X'X)^{-1}X'Y, \tag{2.5.4}$$

is the uniform minimum variance linear unbiased estimator (BLUE) of . Note that for conditions (2.5.3) one must also specify that observations are obtained from a random sample to insure that the Y values are uncorrelated. When one assumes a Gaussian model, i.e., with normally distributed error term, the least squares estimator $\hat{\underline{\beta}}$ is also a maximum likelihood estimator and assumes the additional property of of being a uniform minimum variance unbiased estimator (UMVUE). Graybill (1976) summarizes the relevant statistical facts about the linear regression model and considers the general linear regression model allowing correlated error terms with specified covariance matrix Σ. The general least squares (GLS) estimate of $\underline{\beta}$ is

$$\hat{\underline{\beta}} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \tag{2.5.5}$$

when Σ is known. When the covariance matrix is unknown, it must then be estimated to obtain an estimate for $\underline{\beta}$. Graybill (1976) discusses conditions that the covariance matrix must satisfy to insure that the ordinary least squares (OLS) estimator given by (2.5.4) remains a UMVUE under the more general setting. Estimation of the covariance matrix poses some serious problems to obtaining statistical properties

for GLS estimators of the coefficient vector. Just such a problem occurs in the density estimation procedure discussed in section 4.3. Unfortunately, when the properties of the error vector are only approximately known, one must seek heuristic solutions to the estimation problem.

Parzen (1961) considers a general setting applicable to time series analysis. Essentially, (2.5.1) represents a discrete parameter stochastic process. The continuous parameter analog may be written

$$Y(t) = m(t) + Z(t), \quad t \in T, \tag{2.5.6}$$

where $m(t)$ is a mean value function and $Z(t)$ is a stochastic process with specified properties. One usually assumes that $Z(t)$ has zero mean and covariance kernel

$$K(s,t) = E[Z(s)Z(t)]. \tag{2.5.7}$$

Furthermore, as in the linear regression model, one assumes that $Y(t)$ is observable while $Z(t)$ is not. The following development mirrors Parzen (1961).

Definition 2.5.1 A Hilbert space H with inner product $(\cdot,\cdot)$ is said to be a Reproducing Kernel Hilbert Space (RKHS) with reproducing kernel K if members of H are functions defined on a set T and K is a function on TxT with the following properties for every t in T:

i) $K(\cdot,t)$ as a function of t is in H, and

ii) $(g,K(\cdot,t)) = g(t)$ for every g in H.

Two special stochastic processes are of interest in establishing a parametric theory for continuous parameter regression analysis.

Definition 2.5.2 A stochastic process $\{X(t),\ t\in[0,\infty)\}$ is said to be a Weiner process if $X(0)=0$ and

i) $\{X(t),\ t\in[0,\infty)\}$ has stationary independent increments;

ii) for each $t>0$, $X(t)$ is normally distributed; and

iii) for all $t>0$, $E[X(t)]=0$.

Note that knowledge of the variance of the Weiner process completely characterizes its probability law. For $0\le s\le t$, one observes that

$$\mathrm{Var}[X(t)-X(s)] = \sigma^2(t-s). \tag{2.5.8}$$

A Weiner process is a special case of a normal process. A normal process requires that all finite dimensional distributions be jointly normal. Another special case of a normal process is given by the following definition.

Definition 2.5.3 A stochastic process $\{B(t),\ t\in[0,1]\}$ is called a Brownian Bridge process if it is a normal process with zero mean value function and covariance kernel

$$K(s,t) = \min(s,t) - st. \tag{2.5.9}$$

An important regression model applicable to the density estimation approach of section 4.3 is given by

$$Y(t) = \sum_{k=1}^{m} \theta_k \phi_k(t) + B(t), \quad 0 \le t \le 1, \tag{2.5.10}$$

where $\{\phi_k(t), k=1,m\}$ is a complete system of orthogonal functions in a finite dimensional subspace of $L^2(0,1)$ and $\{B(t), 0 \le t \le 1\}$ is a Brownian Bridge process. Eubank (1979) discusses optimal designs for estimating the $\{\theta_k, k=1,m\}$ based on a finite grid of points in $[0,1]$. One may not have the option to design an experiment to take advantage of these results, however. We shall not consider the problem of optimal designs for such models.

Experience indicates that the OLS estimation techniques often compare favorably in comparison with GLS methods based on estimating an unknown covariance matrix. However, the results are not satisfactory when the discrete covariance structure does not have vanishing off diagonal elements as one moves away from the diagonal. The applications discussed in later chapters avoid such situations.

For the model (2.5.10), Parzen (1961) shows that the maximum likelihood estimates of the parameters are given by

$$\hat{\theta}_k = (Y, \phi_k), \quad j=1,m, \tag{2.5.11}$$

where $(\cdot,\cdot)$ is the inner product of the Hilbert function space

generated by the reproducing kernel K(s,t) of the Brownian Bridge process given in (2.5.9). For nonlinear mean value functions or for infinite expansions utilizing a countable system of basis functions, the problem is more complicated. Desirable properties for the estimates are stated in Parzen (1961), and references are given to proofs from other sources. Specific representations for the estimates will be given in later chapters under nonparametric settings without exploiting the reproducing kernel property. Such generalizations may be desired but will be left to the more mathematically sophisticated researcher.

In some situations, the orthonormal system $\{\phi_k, k=1,m\}$ will be composed of complex valued functions. Brillinger (1975) discusses some generalizations of least squares theory to handle this situation. A complex matrix H is said to be Hermitian if the transpose of H is equal to the conjugate of H. This extends the notion of symmetry for real matrices. One writes H is Hermitian if $H'=\bar{H}$. The definition of nonnegative definiteness readily extends to complex systems. A matrix H is said to be nonnegative definite if

$$\sum_{j=1}^{m}\sum_{k=1}^{m} a_j \bar{a}_k H_{jk} \geq 0 \qquad (2.5.12)$$

for all complex constants $a_1,\ldots,a_m$, where $H=(H_{jk})$ is an m by m complex matrix. The matrices $H'\bar{H}$ and $\bar{H}H'$ are always Hermitian and nonnegative definite.

Now, consider the complex regression model

$$\underline{Y} = H\underline{\beta} + \underline{\varepsilon} \tag{3.4.13}$$

where Y is an nxl observable random vector, the elements of the nxp matrix H are measured without error, $\underline{\beta}$ is a pxl vector of parameters, and $\underline{\varepsilon}$ is an unobservable nxl vector of zero mean uncorrelated random variables with finite variance σ^2. Then

$$(\underline{Y}-H\underline{\beta})'(\overline{\underline{Y}-H\underline{\beta}}) \tag{2.5.14}$$

is minimized over all choices of $\underline{\beta}$ when $\underline{\beta}$ is estimated by

$$\hat{\underline{\beta}} = (H'\overline{H})^{-1}H'\overline{\underline{Y}} \tag{2.5.15}$$

for nonsingular $H'\overline{H}$. Equation (2.5.15) represents the least squares estimates of the parameters in (2.5.13). When the corresponding elements are real, this reduces to the usual least squares formula. Observe that $\underline{Y}$ and $\underline{\varepsilon}$ may be real random vectors and still support complex parameters and design matrix, the only restriction being that complex components of the product must vanish.

3. A SURVEY OF NONPARAMETRIC DENSITY ESTIMATION

3.1 Introduction

In some general cases, parameter estimation is a form of parametric density estimation. Estimating the mean of an exponential distribution or the mean and variance of a normal distribution provides an estimate of the density for that particular distribution. Many goodness-of-fit procedures combine parametric and nonparametric density estimation procedures to arrive at a test statistic for a specified null distribution. The applications of density estimation, however, extend to many areas of statistics. Many parameters of interest are functionals involving the parent density of a data set, so estimating a density can lead indirectly to estimating a parameter of interest. For example, as mentioned in section 2.2, the satistic $\overline{X}_n$ is often written

$$\overline{X}_n = \int_{-\infty}^{\infty} x \, dF_n(x) = (1/n) \sum_{i=1}^{n} X_i, \qquad (3.1.1)$$

where $F_n(x)$ is the empirical c.d.f. based on a random sample of size n, thus emphasizing that $\overline{X}_n$ is an estimator of the parameter

$$\mu = \int_{-\infty}^{\infty} x \, dF(x). \qquad (3.1.2)$$

Writing the above as a Riemann integral, one may wish to form estimates

$$\hat{\mu} = \int_{-\infty}^{\infty} x\, f_n(x)\, dx \qquad (3.1.3)$$

where the integration may be performed numerically. The common grouped data formula for $\bar{X}_n$ given in many elementary statistics texts is actually an integral of a histogram for the data set. Thus, nonparametric density estimation provides a basis for attacking many statistical problems from a nonparametric viewpoint.

Silverman (1980) observes that density estimation is of fundamental importance in exploratory data analysis and has many applications in confirmatory analysis. Silverman notes, however, that "...density estimation cannot be used as a 'back of an envelope' exploratory technique..." like many of the techniques of Tukey (1977), but he does not see this as a disadvantage. In fact, the current state of computer technology and the availability of sophisticated statistical software should make one question any serious exploratory analysis that does not include some form of nonparametric density estimation.

Bean and Tsokos (1980), Tapia and Thompson (1978), Tartar and Kronmal (1976), Carmichael (1976), and Wertz (1980) provide useful bibliographic and expository information regarding the nonparametric estimation of densities and related functions. However, the abundance of literature on the subject should not disguise the fact that the area of nonparametric density estimation is rich with unsolved problems. The fundamental weakness of most procedures is the subjectivity required in choosing a "smoothness parameter". Some objective techniques for handling this problem have been suggested,

but in general the problem presents a serious obstacle to the applicability of most density estimation methods.

This chapter will deal with the major classifications of nonparametric density estimation techniques, giving details of some of the more popular procedures. Comparisons of some of the procedures will be made in Chapter 6, but no Monte Carlo study has been attempted because of the difficulty in handling the subjective smoothing requirements. Anderson (1969, as referenced in Bean and Tsokos, 1980) and Scott and Factor (1981) consider such studies for a restricted set of estimators, but their findings are somewhat inconclusive in terms of the general area of nonparametric density estimation.

Many of the techniques discussed have multivariate extensions. These will be mentioned or referenced , but attention to bivariate density estimation will be withheld until Chapter 4. Both univariate and bivariate techniques will be important in the study of bivariate data analysis. Before considering the techniques, some preliminary concepts need to be introduced.

For the following definitions it will be understood that f_n is an estimate of an unknown p.d.f. f based on a random sample of size n.

Definition 3.1.1 The mean squared error (MSE) of $f_n(x)$ is defined by

$$MSE[f_n(x)] = E\{|f_n(x)-f(x)|^2\} \qquad (3.1.1)$$

If $MSE[f_n(x)] \to 0$ as $n\to\infty$, then one says that $f_n(x)$ is pointwise

consistent in quadratic mean or pointwise consistent in mean square for estimating $f(x)$. If $\sup_x\{MSE[f_n(x)]\} \to 0$ as $n\to\infty$, then one says that f_n is uniformly consistent in quadratic mean for estimating f.

Definition 3.1.2 The mean integrated square error (MISE) of f_n is given by

$$MISE(f_n) = E\{\int_{-\infty}^{\infty}|f_n(x)-f(x)|^2dx\}. \qquad (3.1.2)$$

If $MISE(f_n)\to 0$ as $n\to\infty$, then one says that f_n is uniformly consistent with respect to MISE for estimating f.

The deinitions of weak and strong consistency apply to $f_n(x)$ pointwise, while uniform consistency will imply

$$\sup_x |f_n(x)-f(x)| \quad 0 \qquad (3.1.3)$$

in probability (weakly) or almost surely (strongly).

Two theorems are often exploited regarding applications to density estimation.

Theorem 3.1.1 Let $X_1, X_2, \ldots$, and X be random p-vectors and let $g:R^p \to R$ be a real-valued measurable function that is continuous w.p.1. Then

i) $X_n \xrightarrow{a.s.} X$ implies $g(X_n) \xrightarrow{a.s.} g(X)$;

ii) $X_n \xrightarrow{P} X$ implies $g(X_n) \xrightarrow{P} g(X)$; and
iii) $X_n \xrightarrow{d} X$ implies $g(X_n) \xrightarrow{d} g(X)$.

Serfling (1980, pp. 24-25) proves this theorem and references the more general case where g is vector-valued. By application of this theorem, one obtains the following useful result.

Theorem 3.1.2 Let X_n be $AN(\mu,\sigma_n^2)$ with $\sigma_n^2 \to 0$. Let g be a real-valued measurable function that is differentiable at $x=\mu$, with $g'(\mu) \neq 0$. Then

$$g(X_n) \text{ is } AN(g(\mu), [g'(\mu)]^2\sigma_n^2).$$

As will be seen later, these theorems are usually applied for $g(x)=\log(x)$. Serfling discusses applications to such areas as variance stabilizing transformations and gives several examples utilizing choices for g.

3.2 Nearest Neighbor Density Estimation

Loftsgaarden and Quesenberry (1965) attack the problem of estimating a multivariate density function and arrive at a fairly simple method that possesses desirable properties. Their work appears before that of Cacoullos (1966) who generalized the kernel approach to the multivariate case, and thus represents the first formal development of a technique for multivariate nonparametric density

estimation.

Let $X_1,\ldots,X_n$ be i.i.d. random p-vectors with absolutely continuous c.d.f. $F(x_1,\ldots,x_p)$ and p.d.f. $f(x_1,\ldots,x_p)$. Define

$$V_k(x) = \text{volume of smallest sphere centered at } x \text{ containing at least } k \text{ points.} \quad (3.2.1)$$

Recall, a p-dimensional sphere with radius r has volume

$$V = [\pi^{p/2} r^p]/\Gamma(p/2+1), \quad (3.2.2)$$

where $\Gamma(\cdot)$ is the gamma function. The nearest neighbor density estimate of $f(x)$ is given by

$$f_n(x) = (k/n)/V_k(x), \quad (3.2.3)$$

where $k=k(n)$ is chosen to satisfy certain limiting properties.

Loftsgaarden and Quesenberry (1965) show that $f_n(x)$ is weakly consistent for estimating $f(x)$. Devroye and Wagner (1977) show that with the conditions

$$\begin{aligned} &\text{i) } k(n)\to\infty \text{ and } k(n)/n\to 0; \\ &\text{ii) } k(n)/\log(n)\to\infty; \text{ and} \\ &\text{iii) } f \text{ is uniformly continuous on } R^p, \end{aligned} \quad (3.2.4)$$

then $f_n(x)$ is strongly uniformly consistent for estimating $f(x)$, i.e.,

$$\sup_x |f_n(x) - f(x)| \to 0 \quad \text{a.s.} \tag{3.2.5}$$

For our purposes, we emphasize two theorems due to Moore and Yackel (1977b) and restrict our attention to the univariate case.

Theorem 3.2.1 Let $f(x)$ be given by (3.2.3) and let the following properties hold:

i) f is continuous at x;

ii) $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$; and

iii) $k(n)/\log(\log(n)) \to \infty$.

Then $f_n(x) \to f(x)$ a.s.

Theorem 3.2.2 Let $f_n(x)$ be given by (3.2.3) and let the following properties hold:

i) $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$; and

ii) $\sqrt{k(n)}\,|f_n(x) - f(x)| \to 0$ in probability when $|x_n - x| \le R(n)$,

where $R(n)$ is the radius of the sphere yielding $V_k(x)$. Then

$$\sqrt{k(n)}\,[f_n(x) - f(x)] \xrightarrow{d} N[0, f^2(x)],$$

i.e., $f_n(x)$ is $AN[f(x), f^2(x)/k(n)]$.

As suggested previously, one would prefer that the asymptotic variance did not include an object to be estimated. Theorem 2.3.7 is useful in suggesting variance stabilizing transformations for such cases. Observe, using the log transformation and applying Theorem 2.3.7, one obtains $\log[f_n(x)]$ is $AN[\log f(x), 1/k(n)]$. This result is particularly important to expansion techniques for $\log f(x)$ to be considered in section 4.3.

Moore and Yackel (1977a, 1977b) and Mack and Rosenblatt (1979) consider a more general representation than (3.2.3) and suggest analogies to kernel density estimation. The asymptotic properties of nearest neighbor estimators thus mirror those of kernel density estimators. Kernel estimators are considered in the next section.

3.3 Kernel Density Estimation

The kernel method of density estimation provides a natural extension to the popular histogram estimator and has a firm foundation of approximation theory results to support its use. Rosenblatt (1956) considers this extension of the histogram approach, and Parzen (1962) details the theoretical implications of this technique.

Observe, a histogram estimator of f may be constructed such that the partition of the support of f is composed of equally spaced intervals. Consider the estimator due to Rosenblatt (1956) defined by

$$f_n(x) = \{F_n(x+h) - F_n(x-h)\}/(2h) \qquad (3.3.1)$$

where $h=h(n)$ is a real valued positive function of the sample size n with $h(n)\to 0$ as $n\to\infty$. Rosenblatt shows that for $h(n)=kn^{-\alpha}$, optimal values for k and α may be chosen based on asymptotic mean square error or integrated asymptotic mean square error considerations.

Now, if one lets the kernel K be defined by $K(u)=1/2$ for $|u|<1$ and $K(u)=0$ elsewhere, then (3.3.1) becomes

$$f_n(x) = [1/(nh(n)]\sum_{j=1}^{n} K[(x-X_j)/h(n)]. \qquad (3.3.2)$$

Thus, a more general approach is suggested using the kernel $K(u)$ as a weight function, and forming estimators

$$f_n(x) = \int_{-\infty}^{\infty}[1/h(n)]K[(x-y')/h(n)]\, dF_n(y')$$

$$= [1/nh(n)]\sum_{j=1}^{n} K[(x-X_j)/h(n)]. \qquad (3.3.3)$$

Parzen (1962) gives a table of some common kernels and develops conditions that $K(u)$ must satisfy to obtain desirable statistical properties for the kernel estimator. Two theorems are of importance to us.

Theorem 3.3.1 Let $K(u)$ be a kernel satisfying

i) $\sup_x |K(x)| < \infty$;

ii) $\int_{-\infty}^{\infty} |K(x)|\, dx < \infty$;

iii) $\lim_{x \to \infty} |xK(x)| = 0$; and

iv) $\int_{-\infty}^{\infty} K(x)\, dx = 1$.

Let $f(x)$ be continuous at x, and let $nh(n) \to \infty$ as $n \to \infty$. Then

$$f_n(x) \to f(x) \text{ in q.m.}$$

Theorem 3.3.2 Under the conditions of Theorem 3.3.1, one also has that $f_n(x)$ is $AN[E(f_n(x)), Var\{K[(x-X)/h(n)]\}/(nh^2(n))]$.

Silverman (1978) gives the stronger result of almost sure uniform consistency.

The basic problem one faces when using the kernel method is the choice of window width $h(n)$. Different window widths produce different shapes and in particular introduce the problem of identifying spurious modes. Some degree of subjectivity is hence required to arrive at an acceptable shape for the estimated p.d.f. Silverman (1980) suggests an objective approach to choosing a window width, but the approach necessitates estimation of variance terms whose properties are questionable. Other authors have suggested objective techniques such as cross-validation and "ridge regression" that seem promising but still display weaknesses that cannot be ignored. Nonetheless, kernel density estimation has been extensively studied in the literature and is competitive with other techniques. Cacoullos (1966) extends this technique to the multivariate case.

3.4 Function Approximation Methods

A problem in mathematical analysis is to approximate a member of a suitably restricted class of functions $\{f(x)\}$ by a series expansion technique. The simplest approach seeks coefficients $\{a_j\}_{j=1}^{\infty}$ such that

$$f(x) = \sum_{j=0}^{\infty} a_j x^j \tag{3.4.1}$$

in the sense that

$$\lim_{m\to\infty} |f(x) - \sum_{j=0}^{m} a_j x^j| = 0. \tag{3.4.2}$$

Functions expressable as in (3.4.1) are called <u>entire functions</u> (Davis, 1975). When $f(x)$ is a density, one might consider estimators

$$f_m(x) = \sum_{j=0}^{m} \hat{a}_j x^j \tag{3.4.3}$$

where $\hat{a}_j$ is a function of sample data and m is the order of approximation determined by some meaningful criterion. The usual series expansion for $f(x)$ is a Taylor series, the simplest case given by (3.4.1) with

$$a_j = f^{(j)}(0)/(j!), \quad j=0,1,\ldots, \tag{3.4.4}$$

where $f^{(j)}(0)$ is the j-th derivative of $f(x)$ evaluated at 0.. One

could then base (3.4.3) on estimates of the derivatives of $f(x)$. However, this is usually cumbersome and inefficient in practice.

A more general expansion is given by

$$f(x) = \sum_{j=-\infty}^{\infty} \theta_j \phi_j(x), \quad a \le x \le b \tag{3.4.5}$$

where $\{\theta_k\}_{k=-\infty}^{\infty}$ are real valued constants and $\{\phi_k(x)\}_{k=-\infty}^{\infty}$ is a system of real or complex valued functions. One may then study conditions under which the expansion (3.4.5) is justified. This problem has been studied extensively in the mathematical literature and has recently been applied to statistical problems of density estimation.

Following the development of section 2.4, let H be a separable Hilbert space and let $\{\phi_k(x)\}_{k=-\infty}^{\infty}$ be a complete orthonormal system in H. For most statistical applications, the space of square integrable functions $L^2(a,b)$ is general enough to include a large family of p.d.f.'s and restrictive enough to permit formulation of useful theory; thus, we will henceforth restrict attention to the Hilbert space of square integrable functions. If one chooses an arbitrary member of this space, say $f(x)$, then one is justified in using the expansion (3.4.5) with x in the interval (a,b). Mathematical analysts have studied this expansion for a variety of orthogonal systems of functions. Series expansions of the form (3.4.1) may be translated to (3.4.5) by the Gram-Schmidt orthogonalization of the basis set $\{1,x,x^2,\ldots\}$. Orthogonal polynomials and complex exponentials are two systems that have been studied extensively. The Jacobi polynomials provide a general system of orthogonal polynomials to consider, with

Legendre, Chebyshev, and Laguerre polynomials being special cases (Lanczos, 1956). Complex exponential systems are the basis of Fourier series expansions (Davis, 1975, Churchill, 1969).

For statistical applications, one may seek to estimate a variety of functions using orthogonal expansions. Of primary importance, however, is estimating the c.d.f. or p.d.f. that generates a set of data. One may also estimate the quantile function using these techniques, and some authors base an approach on estimating characteristic functions (see, e.g., Watson and Leadbetter, 1963). We will consider several approaches using orthogonal expansions to estimate an unknown p.d.f. either directly or indirectly. The assumption of a finite support for f is not overly restrictive as one essentially is estimating the truncated density given by

$$f(x) = g(x) / \int_a^b g(x)\,dx, \quad a \le x \le b.$$

Furthermore, transforming a data set to fall in the interval [a,b] is not difficult and has little if any effect on most estimates obtained. Some systems, such as Hermite polynomials, permit expansion over the entire real line, but one is always concerned about extrapolation problems when dealing with a finite data set.

Cencov (1962, taken from Bean and Tsokos, 1980) considers the expansion (3.4.5) where the system $\{\phi_k(x)\}_{k=-\infty}^{\infty}$ is orthogonal w.r.t. a weight function $w(x)$, i.e.,

$$\int_a^b \phi_i(x)\,\phi_j(x)\,w(x)\,dx = \delta(i,j) \qquad (3.4.6)$$

where $\delta(i,j)$ is Kronecker's delta. The coefficients $\{\phi_k\}_{k=-\infty}^{\infty}$ are given by

$$\theta_k = \int_a^b f(x)\,\phi_k(x)\,w(x)\,dx \qquad (3.4.7)$$

and as mentioned in section 2.4, these are usually called Fourier coefficients. Cencov (1962) then obtains estimates for these coefficients based on the empirical c.d.f. From a random sample of size n he obtains

$$\hat{\theta}_k = \int_a^b \phi_k(x)\,w(x)\,dF_n(x) = (1/n)\sum_{j=1}^{n} \phi_k(X_j)\,w(X_j). \qquad (3.4.8)$$

The estimates given by (3.4.8) have some nice properties. Observe,

$$E(\hat{\theta}_k) = (1/n)\sum_{j=1}^{n} E[\phi_k(X_j)\,w(X_j)] = \int_a^b \phi_k(x)\,w(x)\,dF(x) = \theta_k$$

and hence $\hat{\theta}_k$ is unbiased for estimating θ. Furthermore, by the SLLN, $\hat{\theta}_k \to \theta_k$ a.s. This implies that for a finite parameter p.d.f. $f(x)$ with expansion (3.4.5) and $\theta_j = 0$ for $j > m$,

$$f_n(x) = \sum_{k=-m}^{m} \theta_k \phi_k(x), \quad a \le x \le b, \qquad (3.4.9)$$

is unbiased and consistent for estimating $f(x)$. However, $f_n(x)$ will always be biased for estimating infinite parameter models of the form (3.4.5). In this case, the problem then becomes one of choosing the

"best" order m such that $f_n(x)$ in (3.4.9) is a reasonably good estimate of $f(x)$.

One might observe that the Cencov approach is a method of moments estimating scheme. If one rewrites (3.4.7) in terms of expectations, one obtains

$$\theta_k = E[\phi_k(X)w(X)], \tag{3.4.10}$$

and hence (3.4.8) is merely a method of moments estimator of θ_k. As mentioned in Chapter 2, many estimators based upon the e.d.f fall into this category.

Other authors have examined Cencov's technique for specific systems of orthogonal functions. Schwartz (1967) considers expansions based on Hermite polynomials and obtains asymptotic results competitive with kernel estimators. Walter (1977) obtains further results based on this technique. Kronmal and Tartar (1968,1976) consider trigonometric systems, and Crain (1974) uses Legendre polynomials. Anderson (1969) indicates that the Kronmal-Tartar estimators seem to perform better than the Schwartz estimators based on Monte Carlo studies. This would imply that the choice of orthogonal system is crucial to the estimation procedure. Since the literature abounds with various orthogonal expansion techniques, only a few of the more promising ones will be considered.

Kronmal and Tartar (1968) consider estimation techniques based on Fourier series expansions of the c.d.f. $F(x)$ and the p.d.f. $f(x)$. They consider estimators of the form

$$\hat{f}_m(x) = \sum_{j=-m}^{m} a_j \phi_j(x), \quad a \le x \le b \tag{3.4.11}$$

and

$$\hat{F}_m(x) = \sum_{j=-m}^{m} A_j \phi_j(x), \quad a \le x \le b \tag{3.4.12}$$

where $\{\phi_j\}_{j=-m}^{m}$ satisfy (3.4.6). Using trigonometric systems of orthogonal functions they obtain

$$\hat{f}_m(x) = c_0/2 + \sum_{k=1}^{m} c_k \cos[k\pi(x-a)/(b-a)] \tag{3.4.13}$$

where

$$c_k = 2/[(b-a)n] \sum_{i=1}^{n} \cos[k\ (X_i-a)/(b-a)]\ I_A(X_i), \tag{3.4.14}$$

where $A=[a,b]$. This estimate is a derivative of an estimate of $F(x)$ obtained by the expansion (3.4.12) using Cencov type estimates of the coefficients, i.e., using Fourier coefficients based on the e.d.f. Modifications are also suggested to ensure that the density estimate is positive. Tartar and Kronmal (1976) also consider a similar approach based on complex exponentials (i.e., the complex form of the trigonometric systems). From model (3.4.11) they obtain the Cencov type estimates

$$\hat{a}_k = (1/n) \sum_{j=1}^{n} \exp(-2\pi i k X_j), \quad k=-m,\ldots,m, \tag{3.4.15}$$

for the Fourier coefficients. Using an analogy to stepwise regression and the MISE criterion, they omit coefficients $\hat{a}_k$ and $\hat{a}_{-k}$ if

$$\hat{a}_k \hat{a}_{-k} \leq 2/(n+1) \qquad (3.4.16)$$

and terminate the order of approximation when K consecutive coefficients are deemed not significantly different from zero. For practical applications, Tartar and Kronmal suggest letting the maximum order of expansion be m=10, and suggest that K=1 or 2. Large values of m or K usually produce very wiggly estimates of the p.d.f. The MISE stopping rule suggested by Tartar and Kronmal is similar to Parzen's CAT criterion except it emphasizes the contribution of parameters whereas the CAT criterion emphasizes the reduction in residual variance. Such stopping rules add a degree of objectivity to an otherwise subjective endeavor, but detractors often question the ad hoc nature of the criterion functions.

Crain (1974) uses Legendre polynomials as an orthogonal basis set, but he chooses to expand log f(x) instead of f(x) or F(x). Let f(x) be continuous and strictly positive definite such that

$$\log f(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x) - C(\underline{\theta}), \quad a \leq x \leq b, \qquad (3.4.17)$$

where $\{\phi_k(x)\}_{k=1}^{\infty}$ are the Legendre polynomials over [-1,1] and $C(\underline{\theta})$ is an integrating factor insuring that f(x) integrates to one. Consider order m approximators

$$f_m(x) = \exp\{\sum_{k=1}^{m} \theta_k^* \phi_k(x) - (\underline{\theta}^*)\}, a \le x \le b, \qquad (3.4.18)$$

where m and $\{\theta_k^*\}$ are determined by some suitable criterion. Crain uses the criterion of maximum likelihood and establishes conditions that ensure a unique solution vector $\underline{\theta}^*$ exists for the representation (3.4.18). One observes that (3.4.18) is the canonical exponential model representation of a density belonging to a finite parameter exponential family. Furthermore, the expansion of log f(x) rather than f(x) insures that $f_m(x)$ will be positive. One may then treat $\exp\{- C(\underline{\theta}^*)\}$ as an integrating factor to insure that $f_m(x)$ numerically integrates to one.

Sillitto (1969) uses Legendre polynomials shifted to [0,1] to expand the quantile function in a Fourier series and suggests using linear combinations of order statistics to obtain estimates of parameters. Let $X_{1n}, X_{2n}, \ldots, X_{nn}$ be the order statistics from a random sample of size n with strictly increasing (absolutely continuous) c.d.f. F(x). Let $\xi_{pn} = E(X_{pn})$ be the expectation of the p-th order statistic in a sample of size n. Then

$$Q(u) = \sum_{j=1}^{\infty} (2j-1) \lambda_j P_{j-1}^*(u) \qquad (3.4.19)$$

where

$$\lambda_j = (1/j) \sum_{k=0}^{j-1} (-1)^k \binom{j-1}{k} \xi_{j-k,j} \qquad (3.4.20)$$

and $P_{j-1}^*(u)$ is the shifted Legendre polynomial of degree j-1. A

natural estimator of the $\{\lambda_j\}$ is provided by

$$\hat{\lambda}_j = (1/j)\sum_{k=0}^{j-1}(-1)^k\binom{j-1}{k}X_{j-k,j} \qquad (3.4.21)$$

where $X_{j-k,j}=X_{j-k,n}$ (i.e., treating the first j order statistics as if from a sample of size j). Thus, $\hat{\lambda}_j$ is a linear combination of order statistics whose properties are discussed in Sillitto (1969). From the $\{\hat{\lambda}_j\}$ one obtains

$$\hat{Q}(u) = \sum_{j=1}^{n}(2j-1)\hat{\lambda}_j P^*_{j-1}(u). \qquad (3.4.22)$$

To obtain a density estimator, first compute

$$\hat{q}(u) = n\{\hat{Q}[(j+.5)/(n+1)]-\hat{Q}[(j-.5)/(n+1)]\},$$

$$(j-.5)/(n+1) \le u < (j+.5)/(n+1), \qquad (3.4.23)$$

as a raw derivative of $\hat{Q}(u)$ and then use the reciprocal identity [equation (2.2.4)] to obtain

$$\hat{f}(\hat{Q}(u)) = 1/\hat{q}(u), \qquad (3.4.24)$$

which can be plotted for $x=\hat{Q}(u)$ abscissa values to look like a density rather than a density-quantile function if desired.

The nature of functional approximation techniques can lead to a variety of solutions based on the nature of the expansion and the

estimation criterion used. An annoyance is the necessity of considering estimates only in the interval [a,b], but for most applications this poses no real problem. One may wish to investigate which expansions are optimal for specified distributions possessing properties such as symmetry, skewness, wide tails, etc. However, this may be a difficult task with little reward as suggested by some of the simulation studies that have already been performed. Since the primary goal is to estimate an unknown density, one should seek a procedure that prforms well for a large variety of probability models. The class $L^2(a,b)$ provides such a large collection of interesting models, and hence the techniques developed in this section should be competitive for a wide range of parent distributions. An extension of some of the techniques of this section will be considered in section 4.3. For the basic asymptotic results of any particular density estimator discussed in this section, one is referred to the citation corresponding to that procedure. We will have little use of these results for the applications and extensions to be considered later.

3.5 The Autoregressive Approach

The autoregressive approach to density estimation due to Carmichael (1976) and Parzen (1979b) is based upon an analogy between the spectral density of an autoregressive time series and the probability density of a random variable. A density $f(u)$, $0 \le u \le 1$, is said to have an autoregressive representation of order m if it is of the form

$$f(u) = K \left| \sum_{j=0}^{m} a(j) \exp(2\pi i j u) \right|^{-2}, \qquad (3.5.1)$$

where $a(0)=1$, m is a positive integer, K is a positive constant, and $a(1),\ldots,a(m)$ are complex valued coefficients satisfying

$$g(z) = 1 + a(1)z + \ldots + a(m)z \qquad (3.5.2)$$

has all of its roots outside the unit circle. Parzen (1979b) considers the autoregressive representation of the density-quantile function $fQ(u)$.

Analogous to parameter estimation for an autoregressive time series, one estimates the parameters $a(1),\ldots,a(m)$ via the Yule-Walker equations

$$\begin{vmatrix} R(0) & R(1) & \ldots & R(m-1) \\ R(-1) & R(0) & \ldots & R(m-2) \\ \vdots & & & \\ R(1-m) & \ldots & & R(0) \end{vmatrix} \begin{vmatrix} a(1) \\ a(2) \\ \vdots \\ a(m) \end{vmatrix} = - \begin{vmatrix} R(-1) \\ R(-2) \\ \vdots \\ R(-m) \end{vmatrix} \qquad (3.5.3)$$

where $R(v)$ is the Fourier-Stieltjes transform of $F(x)$,

$$R(v) = \int_0^1 \exp(2\pi i v x)\, dF(x), \quad |v| = 0,1,2,\ldots. \qquad (3.5.4)$$

One estimates $R(v)$ by

$$\hat{R}(v) = \int_0^1 \exp(2\pi i v x)\, dF_n(x) \qquad (3.5.5)$$

and obtains $a(1),\ldots,a(m)$ by solving (3.5.3). One also obtains

$$\hat{K}_m = \sum_{j=0}^{m} \hat{a}(j)\hat{R}(j), \tag{3.5.6}$$

where one takes $\hat{a}(0)=1$. The constant K_m becomes an integrating factor, but it corresponds to the prediction variance of an autoregressive time series. In the case of density estimation, K_m will be interpreted as a residual variance to facilitate an objective procedure for determining the best approximating order m. One selects order m such that Parzen's criterion autoregressive transfer function, given by

$$CAT(m) = (1/n)\sum_{j=1}^{m} \hat{K}_j^{-1} - \hat{K}_m^{-1}, \tag{3.5.7}$$

achieves its minimum at m.

Carmichael (1976) gives conditions for the convergence of

$$\hat{f}_m(x) = \hat{K}_m \left| \sum_{j=0}^{m} \hat{a}(j)\exp(2\pi i j x) \right|^{-2} \tag{3.5.8}$$

to the true density $f(x)$. He also relates the autoregressive representation to an approximation in a reproducing kernel Hilbert space using eigenfunctions and eigenvalues corresponding to the reproducing kernel $R(v)$. For more insight into this interpretation, see Parzen (1959, 1967) and Bochner (1955).

Parzen (1979b) develops a goodness-of-fit procedure using the autoregressive technique on a uniform density $d(u)$. First, observe

that under H_0: $fQ(u) = f_0Q_0(u)$ for some specified $f_0Q_0(u)$, the density $d(u)$ defined by

$$d(u) = f_0Q_0(u)/fQ(u), \quad 0 \le u \le 1, \tag{3.5.9}$$

is a uniform density over $[0,1]$. One can then develop a goodness-of-fit procedure based on the sample uniform density defined by

$$\hat{d}(u) = f_0Q_0(u)/\hat{f}\hat{Q}(u), \quad 0 \le u \le 1, \tag{3.5.10}$$

for some estimate $\hat{f}\hat{Q}(u)$ and null value $f_0Q_0(u)$. Parzen develops an autoregressive estimator

$$\hat{d}_m(u) = \hat{K}_m \Big| \sum_{j=0}^{m} \hat{a}(j) \exp(2\pi i j u) \Big|^{-2}, \quad 0 \le u \le 1, \tag{3.5.11}$$

where

$$\hat{K}_m = \int_0^1 \Big| \sum_{j=0}^{m} \hat{a}(j) \exp(2\pi i j u) \Big|^2 \tilde{d}(u)\, du, \tag{3.5.12}$$

$$\tilde{d}(u) = f_0Q_0(u)\tilde{q}(u)/\tilde{\sigma}_0, \tag{3.5.13}$$

with $\tilde{\sigma}_0$ serving as an integrating factor and $\tilde{q}(u)$ representing the empirical quantile-density given by equation (2.2.41). The values $\hat{a}(1),\ldots,\hat{a}(m)$ are derived from the Yule-Walker equations using

$$\tilde{D}(u) = \int_0^{u\sim} d(t)\, dt, \quad 0 \le u \le 1, \qquad (3.5.14)$$

in place of $F_n(x)$ in equation (3.5.5). One may then use $\hat{d}_m(u)$ to form test statistics for testing $H : fQ(u) = f_0Q_0(u)$ against specified alternatives. The estimate $\hat{d}_m(u)$ leads to an estimate of $fQ(u)$ given by

$$\hat{\hat{f}}Q(u) = f_0Q_0(u)/\hat{d}_m(u), \quad 0 \le u \le 1, \qquad (3.5.15)$$

which is based on the representation (3.5.10). Observe that this estimate of $fQ(u)$ is "weighted" by the null density-quantile $f_0Q_0(u)$. Parzen (1979b) suggests that using the normal density-quantile for $f_0Q_0(u)$ provides an essentially nonparametric procedure in that a variety of distributional shapes may still be discovered using this symmetric "smoothing" function.

One of the drawbacks to the autoregressive approach is the difficulty in justifying its use in an intuitive fashion to persons ignorant of autoregressive time series modeling. However, as Parzen (1979b) observes, the knowledge of time series analysis is not essential for one to be able to apply the procedure. The autoregressive approach also seems to be a monster of computational complexity, but many of the computational problems have been overcome by numerical analysts.

There are many advantages to the autoregressive approach to density estimation:

1) It provides an objective means of determining the amount of smoothing required.
2) It provides an abundance of goodness-of-fit diagnostics for a specified null distribution.
3) It has desirable asymptotic properties and seems to perform well for small samples.
4) Computer software is available implementing the procedure (Parzen and Anderson, 1980).

The objective determination of the smoothing order is further enhanced in that it is intuitively justified by the autoregressive model interpretation of the CAT function. One disadvantage to the autoregressive approach is that it may not be extendable to the multivariate case. With this in mind a comparable procedure is developed in the next chapter that readily extends to the bivariate case.

3.6 Other Approaches

In this section we briefly mention techniques that in some cases are variants of the three previous techniques mentioned.

The spline method may be considered as an extension of the kernel method with additional restrictions made to determine the type of smoothing desired and the class of spline functions to be employed. Wahba (1971) considers smoothing the empirical c.d.f. or the empirical quantile function and then differentiates the smoothed estimators to

obtain an estimator for the density. A selling feature of this technique is improved rates of convergence in mean square over the kernel method.

The technique of discrete maximum penalized likelihood (DMPL) estimation as presented by Tapia and Thompson (1978) uses an approach that is a combination of kernel and spline methodologies employing a discrete approximation to a likelihood functional. The resulting estimator is a maximum likelihood estimator (m.l.e.) of a criterion function with an arbitrary smoothness parameter. The object to be maximized is a discrete approximation to the functional

$$L(f) = \prod_{i=1}^{n} f(X_i) \exp\{-\alpha \int_{-\infty}^{\infty} [f''(t)]^2 dt\}. \qquad (3.6.1)$$

Tapia and Thompson present results for this approach along with suggestions for multivariate extensions.

The reciprocal identity employed by Sillitto in section 3.4 is also the basis for the estimator proposed by Bloch and Gastwirth (1968). Their estimate is simply the reciprocal of a raw estimator of $q(u)=Q'(u)$ similar to equation (2.2.41). Their goal concerns asymptotic variance estimation for sample quantiles, and hence they are concerned with pointwise estimation rather than evaluating shapes.

There are many techniques for nonparametric density estimation with each attempting to display some statistical or computational advantage. The references mentioned in the first section of this chapter discuss most of the existing techniques and provide a more comprehensive exposition than contained in this section. Our goal has

been to outline the major classifications of density estimation procedures so as to provide a framework in which to make comparisons with a new technique to be developed in the next chapter. Some comments along these lines are offered in the next section.

3.7 Concluding Remarks

Appraising nonparametric density estimation techniques involves consideration of estimation criterion, robustness, small sample performance, and the nature of the statistical problem of interest. Some techniques may be exceptional for pointwise approximation of a density but lacking when shapes, tail areas,etc., are important. For example, DMPL estimation seems to provide good estimates at a grid of mesh points but somewhat artificially provides the shape of a density. Specifying more mesh points increases computational problems and slows convergence of the algorithm. The emphasis on robustness may hinder evaluating the nature of the tails of a density. Small sample properties may appear satisfactory in simulations, but the problem remains that often small samples do not contain enough information to diagnose weaknesses in the estimate obtained.

Perhaps the most critical problem is the existence of smoothing parameters or orders that must be dealt with in a subjective fashion. The autoregressive technique and the Tartar-Kronmal orthogonal expansion technique suggest criteria for obtaining optimal orders, but further research is warranted into the development of meaningful order determining criteria. However, one may question whether model

selection should be made completely automatic, as such an approach might prevent examining interesting models that may have more theoretical motivation. Automation of a technique could destroy its usefulness in an exploratory analysis.

Another consideration is computational efficiency in light of asymptotic requirements placed on smoothing parameters. One may find it difficult to translate asymptotic restrictions into computer code. Often an upper (or lower) bound is programmed into a procedure so that asymptotic conditions cannot be made to hold, but it would be rather foolish to pay too much attention to this matter since very large data sets may be a rarity. Tartar and Kronmal (1976) suggest that a maximum order of 10 will be adequate for most data sets encountered. The BISAM program discussed in Chapter 5 currently restricts one to a maximum order of 7 which seems adequate for most data sets. The legitimacy of such program restrictions is illustrated by Table 1 which shows some values for common bin width and order parameters as a function of sample size to accomodate the asymptotic theory.

As a final note, the observation is made that much of the literature emphasizes asymptotic properties paying little attention to the practicality of a procedure. While asymptotic properties are desirable, they are worthless when an unmanageable algorithm is required to perform the necessary computations. Unfortunately, this attitude may be carried to extremes as indicated by the overwhelming popularity of histograms. One naturally attempts to seek a balance between theory and computational efficiency. This philosophy is exemplified in the methodology and computer software development

Table 1. Asymptotic Smoothing Orders as a Function of Sample Size

n	log(n)	log(log(n))	SQRT(n)	n**(1/3)
20	3.00	1.10	4.47	2.71
50	3.91	1.36	7.07	3.68
100	4.61	1.53	10.00	4.64
500	6.21	1.83	22.36	7.94
1000	6.91	1.93	31.62	10.00
10000	9.21	2.22	100.00	21.54
100000	11.51	2.44	316.23	46.42
1000000	13.82	2.63	1000.00	100.00

described in the following chapters.

4. BIVARIATE STATISTICAL DATA MODELING

4.1 Introduction

Many of the density estimation techniques of Chapter 3 had multivariate extensions. Bivariate density estimation will provide the framework for the methods of bivariate data analysis that will be developed in later sections. The usual problems of multivariate analysis, however, will present obstacles to direct extension of univariate techniques. One has difficulty in ordering vectors in higher dimensional spaces as well as defining multivariate counterparts to univariate functions. Estimating derivatives of empirical distribution functions is made more difficult and any smoothing must be accomplished for several dimensions. Graphical displays must be broken into component parts for more than three dimensions. Critical regions are more difficult to derive and power considerations for tests of hypotheses may be theoretically impossible.

Our emphasis has been on function estimation and graphical display. For a multivariate problem of more than three dimensions, one may seek to break up the problem into components involving three dimensions or less. As an analogy, recall that the analysis of variance may be treated as multiple t-tests for pairwise comparisons. Naturally, one would only recommend such an attack if the higher dimensional problem had no solution or was too difficult to implement, which is not the case in the analysis of variance analogy. Where

multivariate extensions are not possible, we will recommend treating a multivariate problem as a composition of bivariate problems to be handled by techniques developed in this chapter.

4.2 Normal Theory

As previously mentioned, the usual first step in testing a procedure is to see how it compares to the normal theoretic techniques when data is generated by a normal probability mechanism. In this section, existing normal theoretic results will be examined. For references, Rao (1973), Kshirsagar (1972), or Graybill (1976) provide basic information on existing normal theoretic methods.

Recall, a random p-vector $\underline{X} = (X_1, X_2, \ldots, X_p)'$ has a multivariate normal distribution if its p.d.f. is given by

$$f_{\underline{X}}(\underline{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\underline{X}-\underline{\mu})' \Sigma^{-1} (\underline{X}-\underline{\mu})\} \qquad (4.2.1)$$

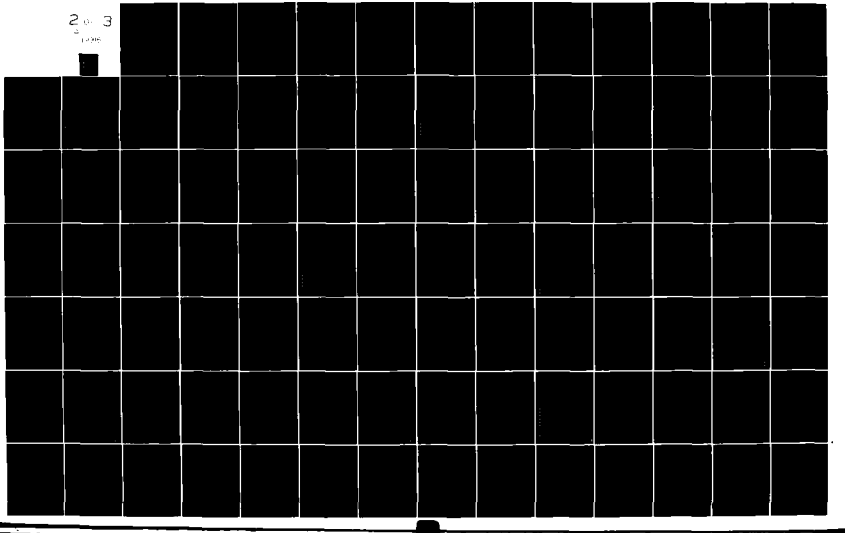
where $\underline{\mu} = E(\underline{X}) = (E(X_1), \ldots, E(X_p))'$ and $\Sigma = (\mathrm{Cov}(X_i, X_j))$. The case p=2 reduces to

$$f_{X,Y}(x,y) = (2\pi)^{-1} (\sigma_X \sigma_Y \sqrt{1-\rho^2})^{-1} \exp\{-1/[2(1-\rho^2)]$$

$$[(x-\mu_X)^2/\sigma_X^2 + (y-\mu_Y)^2/\sigma_Y^2 - 2\rho(x-\mu_X)(y-\mu_Y)/(\sigma_X\sigma_Y)] \qquad (4.2.2)$$

where $-\infty < \mu_X, \mu_Y < \infty$, $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$, and $-1 < \rho < 1$. If $\rho = 0$, one can write $f_{X,Y}(x,y)$ as the product of $N(\mu, \sigma_X^2)$ and $N(\mu, \sigma_Y^2)$ p.d.f.'s, implying

AD-A119 915 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
STATISTICAL MODELING OF BIVARIATE DATA.(U)
AUG 82 T J WOODFIELD
UNCLASSIFIED TR-8-7 DAAG29-80-C-0070
ARO-16992.13-MA NL



that X and Y are independent if $\rho = 0$. This suggests the following well known theorem.

Theorem 4.2.1 If (X,Y) is a bivariate normal random vector, then the correlation between X and Y is zero if and only if X and Y are independent.

This theorem is the basis for many tests of independence, but often its generalization to a nonparametric setting negates the "only if" part of the theorem.

In the parameterization (4.2.2) one may seek statistical estimators of ρ and examine their properties. These statistics may then be employed in testing procedures to test for independence.

The usual approach is to use Pearson's product moment sample correlation coefficient

$$r = \sum_{k=1}^{n} (X_k-\bar{X})(Y_k-\bar{Y}) \Big/ \sqrt{\sum_{k=1}^{n} (X_k-\bar{X})^2 \sum_{k=1}^{n} (Y_k-\bar{Y})^2} \qquad (4.2.3)$$

which is the maximum likelihood estimate for ρ given a random sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ from a bivariate normal distribution. Under $H_0:\rho=0$, one has

$$r\sqrt{(n-2)/(1-r^2)} \sim t(n-2). \qquad (4.2.4)$$

For testing H_0: $\rho=0$ vs. H_1: $\rho\neq 0$ one rejects H_0 at level α if

$$|r|\sqrt{(n-2)/(1-r^2)} > t(\alpha/2;n-2). \qquad (4.2.5)$$

The usual nonparametric approaches to tests of independence are patterned after these normal concepts, i.e., they seek to estimate a population correlation coefficient nonparametrically and base tests on this estimate. Unfortunately, most nonparametric approaches must assume bivariate normality for this approach to be a legitimate test of independence. Huber (1981) notes that r in (4.2.3) is distribution-free under an assumption less restrictive than independence of the bivariate observation (namely, exchangeability of the joint n-vector of X or Y values is assumed), but he observes that r is not invariant to monotone transformations and is very sensitive to outliers in the data. The approaches considered in the next section attempt to overcome such weaknesses.

4.3 Some Concepts, Measures, and Tests of Independence

The primary reference for this section is Lehmann (1966). A brief discussion of the more common nonparametric tests of independence will be followed by a discussion of some useful concepts and measures related to testing for independence. As usual, tests will be based on a bivariate random sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ with assumptions concerning the bivariate distribution of (X,Y).

The discussion of simple linear rank statistics of the form

$$S_n = \sum_{k=1}^{n} c(Q_k)\,a(R_k) \qquad (4.3.1)$$

where $Q_k = \text{rank}(X_k)$, $R_k = \text{rank}(Y_k)$, and the functions $c(i)$ and $a(i)$ are specified score functions, may be found in Serfling (1980) or Ruymgaart (1974). To test the hypothesis H_0: $\rho=0$ vs. some alternative for a suitably defined correlation coefficient ρ one may consider estimators of the form

$$T_n = \sum_{k=1}^{n} [c(Q_k)-\bar{c}][a(R_k)-\bar{a}] / \sum_{k=1}^{n} [c(Q_k)-\bar{c}]^2 \sum_{k=1}^{n} [a(R_k)-\bar{a}]^2 \quad (4.3.2)$$

where $\bar{c} = (1/n) \sum_{k=1}^{n} c(Q_k)$ and $\bar{a} = (1/n) \sum_{k=1}^{n} a(R_k)$. Observe that T_n depends on the sample only through the simple linear rank statistic S_n. T_n is defined analagously to Pearson's product moment sample correlation coefficient, but it has the important additional feature of being invariant to monotone transformations of the data since it depends only on the ranks of the observations.

An important special case of (4.3.2) is given by Spearman's Rho. Let $a(i)=c(i)=i$ so that

$$S_n = \sum_{k=1}^{n} Q_k R_k. \quad (4.3.3)$$

Using the convention of ordering X values and letting R_k be the rank of Y corresponding to X_k, (4.3.3) becomes

$$S_n = \sum_{k=1}^{n} k R_k. \quad (4.3.4)$$

When X and Y are independent, S_n is $AN(\mu_n, \sigma_n^2)$ where

$$\mu_n = n(n+1)^2/4, \quad \sigma_n^2 = n^2(n^2-1)^2/[144(n-1)]. \tag{4.3.5}$$

Define

$$\rho_n = 1 - 6 \sum_{k=1}^{n} (R_k - Q_k)^2/[n(n^2-1)] \tag{4.3.6}$$

which is the Spearman sample correlation coefficient, often called Spearman's Rho (Spearman, 1904, as referenced in Randles and Wolfe, 1979). The population parameter that ρ_n is estimating is given by

$$\rho = 3\ \mathrm{Cov}[\mathrm{sgn}(X_2 - X_1), \mathrm{sgn}(Y_3 - Y_1)]. \tag{4.3.7}$$

For testing H_0: $\rho=0$ vs. H_1: $\rho\neq 0$, using ρ_n, critical values may be found in Table 10 of Conover (1971).

Another popular nonparametric correlation coefficient is Kendall's Tau (Kendall, 1938) based on the concepts of concordance and discordance.

Definition 4.3.1 Two pairs (X_i, Y_i) and (X_j, Y_j) are concordant if $(X_i - X_j)(Y_i - Y_j) > 0$ and discordant otherwise.

Define

$$\tau = 2\ P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1 \tag{4.3.8}$$

or equivalently

$$\tau = \mathrm{Cov}[\mathrm{sgn}(X_2 - X_1), \mathrm{sgn}(Y_2 - Y_1)]. \qquad (4.3.9)$$

Estimate τ by

$$\tau_n = 2\,U_n - 1 \qquad (4.3.10)$$

where U_n is a U-statistic defined by

$$U_n = [1/\binom{n}{2}] \sum\sum_{i<j} h[(X_i, Y_i), (X_j, Y_j)] \qquad (4.3.11)$$

for the kernel h given by

$$h[(X_i, Y_i), (X_j, Y_j)] = I[(X_i - X_j)(Y_i - Y_j)]$$

$$= \mathrm{sgn}(Q_i - Q_j)\,\mathrm{sgn}(R_i - R_j), \qquad (4.3.12)$$

where $I(\cdot)$ is the indicator function defined to be 1 if the argument is positive 0 otherwise. When (X,Y) is a continuous random vector, the theory of U-statistics yields the limiting null distribution of U_n. When X and Y are independent,

$$U_n / [n(n-1)(2n+5)/18]^{1/2} \xrightarrow{d} N(0,1). \qquad (4.3.13)$$

Conover (1971) describes the usual testing procedure and illustrates computational strategies for employing Kendall's Tau in tests of independence. Table 11 of Conover (1971) gives critical values and a

description of the testing procedure for Kendall's Tau. Lehmann (1966) relates this procedure to the difference sign covariance test and suggests similar tests based on the ideas of concordance and discordance. Hajek and Sidak (1967) show that the projection of into the class of linear rank statistics is equivalent to Spearman's Rho.

Blomqvist (1950) develops a procedure that counts the number of data points lying in quadrants I or III when the origin is taken to be (X_0, Y_0). He considers the specific case where X_0 is the median of the X's in the sample and Y_0 is the median of the Y's. Let

$$q = P[(X - X_0)(Y - Y_0) > 0] - P[(X - X_0)(Y - Y_0) < 0]$$

$$= 2\, P[(X - X_0)(Y - Y_0) > 0] - 1 \qquad (4.3.14)$$

and estimate q by

$$Q = 2\, U_n - 1 \qquad (4.3.15)$$

where

$$U_n = (1/n) \sum_{k=1}^{n} 1[(X_k - X_0)(Y_k - Y_0)]. \qquad (4.3.16)$$

Blomqvist (1950) derives the asymptotic normality of Q for a wide range of underlying distributions.

Hoeffding (1948a) discusses some of the above tests in his

classic U-statistic paper and presents the following alternative test in the sequel to this paper (Hoeffding, 1948b). Let

$$\Delta(F) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[F(x,y) - F(x,\infty)F(\infty,y)]^2 dF(x,y). \qquad (4.3.17)$$

Observe that $\Delta(F) = 0$ is equivalent to X and Y being independent. Thus a test based on an estimator of the functional $\Delta(F)$ would be more general than those described above. Let

$$I_D(Z_1,Z_2,Z_3) = I(Z_2-Z_1) - I(Z_3-Z_1) \qquad (4.3.18)$$

and define the kernel h by

$$h[(x_1,y_1),\ldots,(x_5,y_5)] = (1/4)\ I_D(x_1,x_2,x_3)\ I_D(x_1,x_4,x_5)$$
$$I_D(y_1,y_2,y_3)\ I_D(y_1,y_4,y_5). \qquad (4.3.19)$$

Then

$$U_n = [1/\binom{n}{5}]\ \sum h[(X_{i_1},Y_{i_2}),\ldots,(X_{i_5},Y_{i_5})] \qquad (4.3.20)$$

is a U-statistic that is unbiased for estimating $\Delta(F)$. The theory behind this approach requires only that X and Y be continuous random variables. Hence, this is one of the most general nonparametric tests of independence available. The generalization from a parameter ρ to a functional $\Delta(F)$ is a very important step in deriving new

nonparametric tests of independence that are more powerful than existing procedures. This functional approach will be adopted in the sequel using information functionals relevant to the problem of interest.

Given the disparity between nonparmetric tests of independence, one is naturally concerned with the sensitivity of the various test statistics to specific types of dependence between random variables. Hoeffding's procedure based on $\Delta(F)$ seems to be the most sensitive of the approaches mentioned, but due to computational complexity this procedure has not been widely adopted. Lehmann (1966) introduces some concepts relevant to studying dependence between random variables and relates them to existing testing procedures.

Definition 4.3.2 A pair (X,Y) of random variables is said to be positively quadrant dependent if

$$P(X \leq x, Y \leq y) \geq P(X \leq x) P(Y \leq y) \qquad (4.3.21)$$

for all x,y. If the inequality is reversed in (4.3.21), then X and Y are said to be negatively quadrant dependent (NQD). If the inequality holds for at least one pair (x,y), then X and Y are said to be strictly quadrant dependent.

Lehmann (1966) states several theorems relating to positive quadrant dependence. Some general results are given in the following remarks.

Remark 4.3.1 Any bivariate normal random variables with $\rho > 0$ are PQD and with $\rho < 0$ are NQD.

Remark 4.3.2 If X and Y are PQD, then Spearman's Rho (4.3.7), Kendall's Tau (4.3.8), and Blomqvist's q (4.3.14) are all nonnegative.

Remark 4.3.3 If X and Y are PQD, then $\mathrm{cov}(X,Y) \leq 0$. This generalizes Remark 6.3.1 and relates quadrant dependence to covariance. This fact is based on a result due to Hoeffding which states

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[F(x,y) - F(x,\infty)F(\infty,y)]dxdy. \qquad (4.3.22)$$

Observe that dividing inequality (4.3.21) by the positive quantity $P(X \leq x)$ yields

$$P(Y \leq y \mid X \leq x) \geq P(Y \leq y) \qquad (4.3.23)$$

which implies that a knowledge of X being small increases the probability of Y being small. This property is extended to the concept of regression dependence.

Definition 4.3.3 If (X,Y) is a pair of random variables, then Y is positively regression dependent (PRD) on X if $P(Y \leq y \mid X \geq x)$ is non-decreasing in x. If $P(Y \leq y \mid X = x)$ is non-decreasing in x one says that Y is negatively regression dependent (NRD) on X.

Remark 4.3.4 Regression dependence is asymmetric, i.e., Y is PRD on X does not imply X is PRD on Y. For an example of this asymmetry, see Lehmann (1966, pp.1145-1146).

Remark 4.3.5 If Y is PRD (NRD) on X, then X and Y are PQD (NQD). The converse is not necessarily true. Hence, regression dependence is stricter than quadrant dependence.

An even stricter condition than regression dependence is given by the following definition.

Definition 4.3.4 Two random variables X and Y are said to be positively likelihood ratio dependent if their joint c.d.f. satisfies

$$f(x,y')\,f(x',y) \leq f(x,y)\,f(x',y') \qquad (4.3.24)$$

for all $x < x'$, $y < y'$. If the inequality in (4.3.24) is reversed, X and Y are said to be negatively likelihood ratio dependent.

Remark 4.3.6 Likelihood ratio dependence implies quadrant dependence and is symmetric in X and Y.

Remark 4.3.7 Any bivariate normal random variables with $\rho > 0$ are positively likelihood ratio dependent and with $\rho < 0$ are negatively likelihood ratio dependent.

The concepts of regression dependence and likelihood ratio dependence are primarily employed to verify quadrant dependence. The property of quadrant dependence is one of the weakest conditions of dependence for which the popular nonparametric tests are sensitive. The last part of this section suggests some parameters that seem as general as Hoeffding's $\Delta(F)$ in detecting any form of dependence between two r.v.'s.

Kimeldorf and Sampson (1978) consider a condition known as monotone dependence which requires the existence of a monotone function g for which $P[Y=g(X)]=1$. This condition is very restrictive and implies total predictability of Y from X. A less restrictive measure of monotone correlation is thus proposed.

Definition 4.3.5 The monotone correlation ρ^* between two r.v.'s X and Y is given by

$$\rho^* = \sup\{\rho[f(X),g(Y)]\}, \qquad (4.3.25)$$

where $\rho(X,Y)$ defines the correlation between X and Y and the supremum is taken over all monotone functions f and g for which $0<Var[f(X)]<\infty$ and $0<Var[g(Y)]<\infty$.

One may compare this to the sup correlation introduced by Gebelein (see Kimeldorf and Sampson (1978) for reference) which is equivalent to ρ^* except the supremum is taken over all Borel-measureable functions f and g. These concepts are more mathematical than

statistical and are only applied when the parent distributions are known. Kimeldorf and Sampson (1978) do not suggest any estimators for their correlation parameter nor do they propose any testing procedures utilizing the concepts developed. They do, however, point out the desirable property that $\rho^*=0$ is equivalent to X and Y being independent, and hence they have developed a correlation parameter of special interest in tests of independence. Apparently sup correlation has this same property. One need only find estimators of these parameters to develop a powerful nonparametric test of independence. Clearly this is an awesome task that has yet to be fully implemented.

Hajek and Sidak (1967) identify locally most powerful rank tests (LMPRT) for testing for independence. Hoeffding (1948b) discusses the problem of obtaining unbiased tests against all alternatives. Blum, Kiefer, and Rosenblatt (1961) suggest a competitor to Hoeffding's U-statistic approach based on a statistic analagous to the Cramer-von Mises goodness-of-fit statistic, but their approach is as computationally complex as Hoeffding's approach and has received little attention in statistical applications. Gibbons (1971) and Conover (1971) consider some traditional requirements that a measure of association is expected to satisfy and check these requirements for the popular nonparametric statistics. They fail to suggest updating these requirements to included estimators of functionals like the one proposed by Hoeffding (1948b). Until estimators and testing procedures are developed for the correlation parameters considered by Kimeldorf and Sampson (1978), the use of functionals to measure dependence seems to be the most promising method of developing general

nonparametric procedures for bivariate data analysis.

4.4 Bivariate Density Estimation Using Information Criterion

Let $(X_1,Y_1),\ldots.(X_n,Y_n)$ be a bivariate random sample with joint c.d.f. $F_{X,Y}$ marginals F_X, F_Y, and associated functions under the usual notation. Form the uniform bivariate sample $(Q_1/(n+1),R_1/(n+1)),\ldots,(Q_n/(n+1),R_n/(n+1))$ where $Q_i = \text{rank}(X_i)$ and $R_i = \text{rank}(Y_i)$. One may then treat this as a sample from the dependence density $d(u_1,u_2)$ to form estimates $\tilde{d}(u_1,u_2)$ using generalizations of the techniques of Chapter 3.

The nearest neighbor techniques of section 3.2 requires no generalization since it was designed for multivariate density estimation. The theorems stated hold for the bivariate case. One notes that this technique has an advantage over other bivariate procedures in that subjective considerations of the smoothing parameter $k(n)$ are not unduly complicated by multivariate generalizations. The value $V_k(x)$, however, becomes more complicated in higher dimensional settings. This remains one of the easier computational techniques in the bivariate case when compared with other approaches.

Cacoullos (1966) generalizes kernel density estimation to include multivariate estimators of the form

$$f_n(x_1,\ldots,x_p) = (1/n\prod_{i=1}^{p}h_i)\sum_{j=1}^{n}K[(x_1-X_{j1})/h_1,\ldots,(x_p-X_{jp})/h_p]. \quad (4.4.1)$$

Often one chooses the kernel function to be a product of univariate kernels. Cacoullos (1966) proves multivariate extensions to most of the theorems found in Parzen (1962). In the bivariate case, one has

$$f_n(x,y) = [1/n(h_1h_2)] \sum_{j=1}^{n} K[(x-X_j)/h_1, (y-Y_j)/h_2] \qquad (4.4.2)$$

which is usually taken to be

$$f_n(x,y) = [1/n(h_1h_2)] \sum_{j=1}^{n} K_1[(x-X_j)/h_1] \; K \; [(y-Y_j)/h_2] \qquad (4.4.3)$$

for univariate kernels K_1 and K_2 . For this estimator, the window width problem is essentially raised to a power of 2. For example, looking at estimates for three different window widths in the univariate case would expand to looking at nine different estimators in the bivariate case to include all possible combinations of window widths. Nonetheless, this technique remains one of the more popular methods of bivariate density estimation.

Tartar and Kronmal (1970) consider p-dimensional Fourier expansion methods to obtain some theoretical results for multivariate density estimation. Tartar and Silvers (1975) apply an orthogonal expansion technique to the estimation of a bivariate density and suggest theoretical implications and applications for decomposing a mixture of Gaussian distributions. We propose a new approach that is based on the Tartar and Kronmal estimation scheme except that a different estimation criterion is employed. The approach is motivated by seeking a more sophisticated estimation criterion than the method

of moments and basing it on a "better" initial estimate that the empirical c.d.f. We will introduce this new estimator by first considering the univariate case.

Let $X_1,\ldots,X_n$ be i.i.d. r.v.'s with p.d.f. f. Let log f be in $L^2(a,b)$. If $\{\phi_k(x)\}_{k=-\infty}^{\infty}$ is a complete orthonormal system of functions in $L^2(a.b)$, then

$$\log f(x) = \sum_{k=-\infty}^{\infty} \theta_k \phi_k(x) - C(\underline{\theta}), \quad a \le x \le b, \tag{4.4.4}$$

where $\{\theta_k\}_{k=-\infty}^{\infty}$ are real valued constants and $C(\underline{\theta})$ is an integrating factor to insure that f(x) integrates to one. Following Crain (1974), one may consider order m approximators

$$\log f_m(x) = \sum_{k=-m}^{m} \theta_k \phi_k(x) - C(\underline{\theta}), \quad a \le x \le b, \tag{4.4.5}$$

and attempt to find estimates $\{\hat{\theta}_k\}_{k=-m}^{m}$ that possess desirable statistical properties to yield an estimate of $f_m(x)$.

If one chooses the criterion of minimum information, one seeks parameter estimates that minimize $I(f_m;\hat{f}_m)$ where

$$\log \hat{f}_m(x) = \sum_{k=-m}^{m} \hat{\theta}_k \phi_k(x) - C(\hat{\underline{\theta}}), \quad a \le x \le b, \tag{4.4.6}$$

However, f_m is unrealizable so that the quantity $I(f_m;\hat{f}_m)$ can only be examined from a limiting perspective as in Crain (1974). Furthermore, different choices of m yield different estimators with the ultimate goal being the estimation of the true density f.

To overcome these problems consider an alternate definition of information. Recall the definition of the bi-information between two densities given by (2.3.10). Using this measure of information as an estimation criterion, the problem may be rewritten to resemble an exercise in continuous parameter regression analysis.

Let $\tilde{f}_n$ be a "raw" estimator of f satisfying

i) $\tilde{f}_n \to f$ in q.m. or a.s., and

ii) $\{\tilde{f}_n(x)\}$ is asymptotically a Gaussian process.

Conditions (4.4.7), Theorem 3.1.1, and Theorem 3.1.2 guarantee the appropriate behaviour for $\log \tilde{f}_n$ as required below. The conditions are stated in terms of $\tilde{f}_n$ since this is how they are usually found in the literature. Consider the approximate model

$$\log \tilde{f}_n(x) = \sum_{k=-m}^{m} \theta_k \phi_k(x) + \alpha_n G(x), \quad a \le x \le b, \tag{4.4.8}$$

where $G(x)$ is a Gaussian process and α_n is a "generalized variance". Using bi-information based on empirical measure and representation (4.4.6) for an estimator of f, observe

$$I(\tilde{f}_n ; \hat{f}_m) = \int_a^b \left| \log \tilde{f}_n(x) - \sum_{k=-m}^{m} \hat{\theta}_k \phi_k(x) \right|^2 df_n(x)$$

$$= (1/n) \sum_{i=1}^{n} \left| \log \tilde{f}_n(X_i) - \sum_{k=-m}^{m} \hat{\theta}_k \phi_k(X_i) \right|^2 \tag{4.4.9}$$

The constant term in practice is omitted during the parameter estimation phase and re-introduced later as an integrating factor. Equation (4.4.9) indicates that minimum bi-information estimators are equivalent to least squares estimators. Estimators of the form (4.4.6) may then be easily derived using one's favorite least squares regression computer program. Furthermore, if $\tilde{f}_n(x)$ is chosen so that $\log \tilde{f}_n$ is in $L^2(a,b)$, the approximation theory for Hilbert spaces insures that the least squares estimates of the parameters will be Fourier coefficients for a suitably orthonormalized system of "independent variables".

One still faces the problem of determining the "optimal" order m, but in the regression framework several approaches are suggested. Hocking (1976) considers a variety of stepwise regression techniques that may be useful in selecting a best order m. Time series criterion functions used in determining optimal orders for autoregressive models may also be useful, Parzen's CAT (Criterion Autoregressive Transfer) function and Akaike's information criterion (AIC) function being primary candidates for consideration. The MISE criterion of Tartar and Kronmal (1970,1976) may also be employed.

To emphasize the Hilbert space approach to approximation theory, suppose that the estimator $\log \tilde{f}_n$ is defined so as to be square integrable (which is usually the case since most estimates will be bounded with finite support). Let $\{\phi_k(x)\}_{k=-\infty}^{\infty}$ be an orthogonal system w.r.t. $dF_n(x)$, i.e.,

$$\int_a^b \phi_i(x)\, \phi_j(x)\, dF_n(x) = \delta(i,j), \qquad (4.4.10)$$

where $\delta(i,j)$ is Kronecker's delta. Observe,

$$I(\tilde{f}_n;f_m) = \int_a^b |\log \tilde{f}_n(x) - \sum_{k=-m}^{m} \theta_k \phi_k(x)|^2 dF_n(x)$$

$$= \int_a^b |\log \tilde{f}_n(x)|^2 dF_n(x) + \int_a^b |\sum_{k=-m}^{m} \theta_k \phi_k(x)|^2 dF_n(x)$$

$$-2 \sum_{k=-m}^{m} \theta_k \int_a^b [\log \tilde{f}_n(x)] \phi_k(x) \, dF_n(x).$$

Squaring the appropriate terms and taking advantage of the orthogonality property (4.4.10) one obtains

$$I(\tilde{f}_n;f_m) = (1/n) \sum_{i=1}^{m} \log \tilde{f}_n(X_i) + \sum_{k=-m}^{m} \theta_k^2$$

$$-2 \sum_{k=-m}^{m} \theta_k [(1/n) \sum_{i=1}^{n} \log \tilde{f}_n(X_i) \phi_k(X_i)].$$

Taking derivatives w.r.t. θ_K and setting the equations equal to zero, one has

$$\hat{\theta}_k = (1/n) \sum_{i=1}^{n} \log \tilde{f}_n(X_i) \phi_k(X_i), \qquad (4.4.11)$$

which is the Fourier coefficient of the expansion (4.4.5) w.r.t. empirical measure. One may easily verify that the estimates defined by (4.4.11) indeed minimize $I(\tilde{f}_n;f_m)$ so that a minimum bi-information estimator has been obtained.

For the estimator $\hat{f}_m(x)$ given by

$$\hat{f}_m(x) = \exp\Big[\sum_{k=-m}^{m} \theta_k \phi_k(x)\Big] \qquad (4.4.12)$$

to be consistent, the estimator $\tilde{f}_n(x)$ must be chosen so that

$$[\log \tilde{f}_n(x) - \log \hat{f}_m(x)] = o(1/\sqrt{n}) \text{ a.s.} \qquad (4.4.13)$$

By Minkowski's inequality,

$$\| \log \hat{f}_m(x) - \log f(x) \| = \| \log \hat{f}_m(x) - \log \tilde{f}_n(x) + \log \tilde{f}_n(x) - \log f(x) \|$$

$$\leq \| \log \hat{f}_m(x) - \log \tilde{f}_n(x) \| + \| \log \tilde{f}_n(x) - \log f(x) \|.$$

The second term of the right hand side of the inequality converges almost surely to zero by assumptions (4.4.7), and hence $\hat{f}_m(x)$ converges almost surely to $f(x)$ by assumption (4.4.13) and Theorem 3.1.1. Although this result seems straightforward, one observes that it may be very difficult to verify assumption (4.4.13) for a particular estimator $\tilde{f}_n(x)$ because of the difficulty in understanding the behaviour of the two estimators as both m and n approach infinity.

To show the asymptotic normality of $\hat{f}_m(x)$, let conditions (4.4.7) and (4.4.13) hold with $\log \tilde{f}_n(x)$ being $AN[\log f(x), \sigma_n^2]$. Furthermore, let the asymptotic variance σ_n^2 be independent of $f(x)$ and let $\sigma_n^2 \to 0$ and $n\sigma_n^2 \to \infty$ as $n \to \infty$. This usually follows when $\tilde{f}_n(x)$ is consistent for estimating $f(x)$ and taking the logarithm of $\tilde{f}_n(x)$ is a variance stabilizing transformation. Then

$$[\log \hat{f}_m(x) - \log f(x)]/\sigma_n$$

$$= [\log \hat{f}_m(x) - \log \tilde{f}_n(x)]/\sigma_n + [\log \tilde{f}_n(x) - \log f(x)]/\sigma_n$$

$$= (1/\sigma_n\sqrt{n})\ \sqrt{n}[\log \hat{f}_m(x) - \log \tilde{f}_n(x)] + [\log \tilde{f}_n(x) - \log f(x)]/\sigma_n$$

$$= (1/\sigma_n\sqrt{n})\, A(n) + B(n)\,.$$

By assumption, $(1/\sigma_n\sqrt{n}) \to 0$ as $n \to \infty$, and by application of (4.4.13), $(1/\sigma_n\sqrt{n})\, A(n) \to 0$ in probability as $n \to \infty$. Furthermore, $B(n)$ converges in distribution to a $N(0,1)$ r.v. by choice of initial estimator $\tilde{f}_n(x)$, and hence by Slutsky's Theorem, $\log \hat{f}_m(x)$ is $AN[\log f(x), \sigma_n^2]$.

Condition (4.4.13) severely limits these results and has not been shown to hold for any of the common nonparametric estimators. The nearest neighbor estimate satisfies the asymptotic normality requirements with stabilized variance, but the more stringent condition (4.4.13) has not been verified. Nonetheless, the estimator $\hat{f}_m(x)$ is intuitively appealing as well as being the optimal estimator by use of an information criterion. Applications considered in Chapter 6 will further support the use of this new estimator by considering comparisons with some of the nonparametric density estimators discussed in Chapter 3. However, our main concern is bivariate extensions in the quantile domain to overcome some of the weaknesses of other bivariate density estimators.

The extension of the orthogonal expansion technique to the bivariate case is relatively straightforward. A bivariate orthonormal

system is fairly easy to obtain as the following theorem indicates.

Theorem 4.4.1 Let $\{\phi_k(x)\}_{k=-\infty}^{\infty}$ be a complete orthonormal system for $L^2(a,b)$. Then $\{\phi_j(x)\,\phi_k(y)\}_{j,k=-\infty}^{\infty}$ is a complete orthonormal system for $L^2(A)$ where $A=\{(x,y): a\le x,y\le b\}$.

Proof:
$$\int_a^b\int_a^b [\phi_i(x)\,\phi_j(y)][\phi_k(x)\,\phi_l(y)]\,dx\,dy$$

$$= \int_a^b [\phi_i(x)\,\phi_k(x)]\,dx \int_a^b [\phi_j(y)\,\phi_l(y)]\,dy$$

$$= \begin{cases} 1 & \text{if } i=k \text{ and } j=l \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $\{\phi_j(x)\,\phi_k(y)\}_{j,k=\infty}^{\infty}$ is an orthonormal system for $L^2(A)$. To show that this system is complete, first observe that if $g(x,y)\in L^2(A)$, then for fixed y, treated as a function of x, $g(x,y)\in L^2(a,b)$, and vice versa. Define

$$h(x) = \int_a^b \phi_k(y)\; g(x,y)\; dy.$$

Then

$$\int_a^b\int_a^b \phi_j(x)\phi_k(y)\; g(x,y)\; dx\,dy = 0$$

implies

$$\int_a^b \phi_j(x)\ h(x)\ dx = 0$$

and by completeness of $\{\phi_k(x)\}$ it follows that $h(x)=0$ a.e. Thus, by definition of $h(x)$

$$\int_a^b \phi_k(y)\ g(x,y) = 0$$

which implies for fixed x, $g(x,y)=0$ a.e. Reversing the order of integration one obtains $g(x,y)=0$ a.e. and hence $\{\phi_j(x)\phi_k(y)\}_{j,k=-\infty}^{\infty}$ is a complete orthonormal system for $L^2(A)$.■

This theorem allows us to employ one of the many popular univariate orthonormal systems in the approximation of bivariate densities.

Orthogonal expansion techniques still possess the problem of choosing an order of approximation. However, as with nearest neighbor density estimation, the problem is not as sensitive to dimensionality increases as with the kernel method. Tartar and Kronmal (1970,1976) suggest a stopping rule in sequentially adding terms that is based on the sample mean integrated square error. Such stopping rules are useful but may prevent one from observing interesting shapes that may result from addition of extra terms. Furthermore, some degree of subjectivity is always inherent in order selection criteria despite heuristic motivations.

To use the techniques of section 3.4 applied to estimating $d(u_1,u_2)$, the following theorem is necessary.

Theorem 4.4.2 Let $\log f_X$, $\log f_Y$ be in $L^2(a,b)$ and let $\log f_{X,Y}$ be in $L^2(A)$ where A is defined as in Theorem 4.4.1. If $d(u_1,u_2)$ is defined as in (2.2.17), then $\log d$ is in $L^2(B)$ where B is the unit square, $B = \{(u_1,u_2): 0 \le u_1,u_2 \le 1\}$.

Proof: Observe

$$\int_0^1\int_0^1 |\log d(u_1,u_2)|^2 \, du_1 du_2$$

$$= \int_0^1\int_0^1 |\log[f_{X,Y}(Q_X(u_1),Q_Y(u_2))/f_X(Q_X(u_1))f_Y(Q_Y(u_2))]|^2 du_1 du_2$$

$$= \int_0^1\int_0^1 |\log f_{X,Y}(Q_X(u_1),Q_Y(u_2)) - \log f_X(Q_X(u_1)) - \log f_Y(Q_Y(u_2)|^2 du_1 du_2$$

$$= \int_0^1\int_0^1 |\{|\log f_{X,Y}(Q_X(u_1),Q_Y(u_2))|^2 + |\log f_X(Q_X(u_1))|^2$$

$$+ (\text{cross product terms})\}| \, du_1 du_2$$

$$\le \int_0^1\int_0^1 |\log f_{X,Y}|^2 du_1 du_2 + \int_0^1 |\log f_X|^2 du_1$$

$$+ \int_0^1 |\log f_Y|^2 du_2 + 2\int_0^1\int_0^1 |\log f_{X,Y} \log f_X| du_1 du_2$$

$$+ 2\int_0^1\int_0^1 |\log f_{X,Y} \log f_Y| du_1 du_2 + 2\int_0^1\int_0^1 |\log f_X \log f_Y| du_1 du_2$$

by Minkowski's inequality, where we have adopted the abbreviated notation

$$f_{X,Y} = f_{X,Y}(Q_X(u_1), Q_Y(u_2)),\ f_X = f_X(Q_X(u_1)),\ f_Y = f_Y(Q_Y(u_2)).$$

By assumption, the first three terms on the right hand side of the inequality are fintite. By Holder's inequality,

$$\int_0^1\int_0^1 |\log f_{X,Y} \log f_X| du_1 du_2 \le [\int_0^1\int_0^1 |\log f_{X,Y}|^2 du_1 du_2]^{\frac{1}{2}}$$

$$[\int_0^1 |\log f_X|^2 du_1]^{\frac{1}{2}} < \infty,$$

with the last inequality following by assumption. The finiteness of the remaining terms follows similarly, and hence the theorem is proved.■

The fact that log d is in $L^2(B)$ allows an orthogonal expansion of log d, and hence one can apply the approximation techniques of section 3.4.

The use of complex exponential orthogonal systems in a Fourier expansion will necessitate the use of complex least squares procedures to carry out the *minimum information* approach. Specifically, one considers the expansion

$$\log d(u_1,u_2) = \sum_{j,k=-\infty}^{\infty} \theta_{jk}\phi_j(u)\phi_k(u) - C(\underline{\theta}) \qquad (4.4.14)$$

where the $\{\theta_{jk}\}_{j,k=-\infty}^{\infty}$ are complex valued parameters, $\{\phi_j(u)\}_{j=-\infty}^{\infty}$ is a univariate complex orthogonal system, and $C(\underline{\theta})$ is a complex integrating factor. Since $\log d(u_1,u_2)$ is real, the contribution of

the complex terms must vanish. This will occur if conjugate pairs always appear together. For a finite parameter model of order m, one estimates the $\{\theta_{jk}\}_{j,k=-m}^{m}$ by complex least squares after deriving an initial estimate $\tilde{d}(u_1,u_2)$ of $d(u_1,u_2)$. The minimum bi-information estimate $\hat{d}(u_1,u_2)$ is then obtained from the model

$$\log \hat{d}(u_1,u_2) = \sum_{j,k=-m}^{m} \hat{\theta}_{jk}\phi_j(u_1)\phi_k(u_2) - C(\underline{\hat{\theta}}) \qquad (4.4.15)$$

where $C(\underline{\hat{\theta}})$ is chosen so that $d(u_1,u_2)$ integrates to one.

Some consequences of this approach are worth noting. If the initial estimate $\tilde{d}(u_1,u_2)$ is also derived using only the rank transformations of the data, then $\hat{d}(u_1,u_2)$ is a fully nonparametric estimator requiring only an assumption of continuous data and square integrability of the logarithms of the underlying joint and marginal p.d.f.'s. Furthermore, $\hat{d}(u_1,u_2)$ is invariant to monotone transformations of the data since it is a ranking procedure. The parametric representation of $\hat{d}(u_1,u_2)$ permits complete specification of the model by only knowing the values of m^2 estimates of the parameters unlike the nearest neighbor, kernel, and penalized likelihood approaches. The problem of defining the region of support is surmounted by the uniform transformations to the unit square. The estimate also possesses many of the desirable properties of its univariate counterpart although asymptotic properties are confounded further by approximating a uniform data set by rank transformed data.

The problem of smoothing or order determination has many heuristic rules of thumb all of which need further research. The CAT

and AIC criterion functions seem to recommend too many parameters in initial investigations done by the author, while a minimum information (maximum entropy) criterion seems to pick too few parameters and hence produces an overly smoothed estimate. As mentioned earlier, Tartar and Kronmal (1970,1976) suggest a minimum MISE criterion that picks the smallest order m such that

$$\theta_{m,m} \bar{\theta}_{-m,-m} > (n+1)^{-1}. \qquad (4.4.16)$$

One is naturally concerned that the inclusion of too many parameters will introduce spurious modes, so one recommendation is to produce three estimates with varying degrees of smoothness. One may then hope that the *physical constraints* of the problem or the expertise of the experimenter will aid in model selection.

Interpreting three dimensional graphs and contour plots of estimates of the dependence density is particularly difficult, due in part to the radical nature of this approach to data analysis. Consequently, one may prefer to form estimates of the bivariate density-quantile function $fQ(u_1,u_2) = f(Q_X(u_1),Q_Y(u_2))$. The approach we favor forms

$$\hat{\hat{fQ}}(u_1,u_2) = \hat{d}(u_1,u_2)\ \hat{\hat{fQ}}_X(u_1)\hat{\hat{fQ}}_Y(u_2) \qquad (4.4.17)$$

where $\hat{d}(u_1,u_2)$ is a minimum bi-information estimator of $d(u_1,u_2)$ and the estimated density-quantiles are obtained using the autoregressive method. This approach allows one to take advantage of the

autoregressive approach to univariate data modeling. Goodness-of-fit tests may be conducted for null distributions of the univariate densities.

Experience with this approach reveals several interesting features that are of importance in bivariate data analysis. For local alternatives to independence, in particular, for cases when the linear correlation exists and is "small", the univariate density-quantiles dominate the shaping of the bivariate density-quantile. If either univariate density is bimodal and the correlation is small, $\hat{d}(u_1,u_2)$ will closely approximate a flat surface so that the influence of the bimodal univariate density will create a bimodal or multimodal bivariate density-quantile. However, when two random variables are highly correlated, the dependence density dominates the shaping of the bivariate density-quantile and tends to smooth out any anomalies in the univariate density-quantiles.

In practical applications one is particularly concerned that a univariate density estimation technique not introduce modes that will unduly affect the bivariate density-quantile function. An example in Chapter 6 illustrates a situation where an outlier in a data set introduces a spurious mode in the univariate density estimate of one of the variables thereby causing the bivariate fQ function to be multimodal. The estimate $\hat{d}(u_1,u_2)$ is unaffected by the outlier, but the autoregressive approach is unduly influenced. For this example, the outlier was easily detected so that it could be removed, but one remains concerned about the sensitivity of the autoregressive estimate. In particular, one is interested in the ability of the AR

approach to detect unimodal and bimodal shapes, since clearly outliers that cannot be explained by measurement error usually suggest bimodality. A study was performed for 50 iterations of three types of samples of size 100. The first sample represents a N(0,1) distribution, the second represents a sample from the mixture 0.5 N(0,1) + 0.5 N(2,2), and the third sample comes from the mixture 0.5 N(0,1) + 0.5 N(3,2). The results are given in Table 2. Order 0 values indicate "acceptance" of a null hypothesis of normality. Order

Table 2. Monte Carlo Study of CAT Criterion for Density Estimation

Sample	Order	Frequency
N(0,1)	0	42
	1	7
	2	0
	3	1
	4	0
.5 N(0,1) + .5 N(2,2)	0	26
	1	19
	2	1
	[illegible]	3
	[illegible]	1
.5 N(0,1) + .5 N(3,2)	0	5
	1	38
	2	5
	3	1
	4	1

1 values suggest slight skewness and hints of bimodality. Order 2 values suggest bimodality with possible hints of trimodality, etc.

When the modes are distinct as in the third case, the AR modeling approach performs well, whereas in case 2 the technique found it difficult to distinguish the modes. Note that high order selections are very rare for "smooth" parent densities. One interpretation of the selection of a high order is that outliers may be present in the data, which is useful for the application of bivariate density estimation to data analysis.

In Chapter 6 we will illustrate the use of the estimated bivariate density-quantile to locate modes in a bivariate distribution. Naturally, one may wish to investigate this approach with different estimators for the univariate density-quantile functions, but as stated, we feel that the AR approach has the most objective and consistent results.

4.5 Some Entropy-Based Measures of Association

In section 4.2 some popular nonparametric tests of independence were considered and viewed in light of some concepts and measures of dependence. The observation was made that only certain functionals and variations of sup correlation were general enough to detect all deviations from independence. In this section a new functional based on the concept of information is introduced that is as general as Hoeffding's $\Delta(F)$, and various testing procedures are proposed using this new measure of dependence.

Let (X,Y) be bivariate random variables with joint c.d.f. $F_{X,Y}$, joint p.d.f. $f_{X,Y}$, marginal c.d.f.'s F_X and F_Y, marginal p.d.f.'s f_X

and f_Y, and quantile functions Q_X and Q_Y. The dependence distribution function $D(u_1,u_2)$ and the dependence density $d(u_1,u_2)$ are defined as in section 2.2, equations (2.2.16) and (2.2.17). Using equation (2.3.1) defining information, one obtains the information between the joint p.d.f. $f_{X,Y}$ and the product of the marginals f_X and f_Y by

$$I(f_{X,Y};f_X f_Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\log[f_{X,Y}(x,y)/f_X(x)\,f_Y(y)]\} \cdot f_{X,Y}(x,y)\,dx\,dy. \quad (4.5.1)$$

With the usual change of variable $u_1 = F_X(x)$ and $u_2 = F_Y(y)$, one obtains

$$I(f_{X,Y};f_X f_Y) = \int_0^1\int_0^1 \log[f_{X,Y}(Q_X(u_1),Q_Y(u_2))/f_X(Q_X(u_1))\,f_Y(Q_Y(u_2))] \cdot f_{X,Y}(Q_X(u_1),Q_Y(u_2))\; q_X(u_1)q_Y(u_2)\;du_1du_2 \quad (4.5.2)$$

which reduces to

$$I(f_{X,Y};f_X f_Y) = \int_0^1\int_0^1 [\log d(u_1,u_2)]\; d(u_1,u_2)\;du_1du_2 = -H(d), \quad (4.5.3)$$

where $H(d)$ is the entropy of the dependence density. From the information inequality one obtains

$$I(f_{X,Y}; f_X f_Y) = 0 \quad \text{iff} \quad f_{X,Y}(x,y) = f_X(x) f_Y(y), \text{ a.e.,} \qquad (4.5.4)$$

and thus equation (4.5.3) leads to new techniques for ascertaining whether X and Y are independent.

The technique to be investigated will estimate H(d) by estimating d and using various numerical or statistical integration procedures. Then Monte Carlo studies will be employed to investigate the properties of the estimator.

One solution is to estimate d by $\tilde{d}$ and then form

$$H(\tilde{d}) = \int_0^1 \int_0^1 [\log \tilde{d}(u_1, u_2)] \, d\tilde{D}(u_1, u_2)$$

$$= (1/n) \sum_{k=1}^{n} \log \tilde{d}(Q_k/(n+1), R_k/(n+1)) \qquad (4.5.5)$$

where $\tilde{D}(u_1, u_2)$ is the empirical dependence distribution function with jumps of size 1/n at the points $(Q_k/(n+1), R_k/(n+1))$. Recall, $Q_k = \text{rank}(X_k)$ and $R_k = \text{rank}(Y_k)$. Another solution might be to numerically integrate

$$H(\tilde{d}) = \int_0^1 \int_0^1 \{\log \tilde{d}(u_1, u_2)\} \, \tilde{d}(u_1, u_2) \, du_1 du_2 \qquad (4.5.6)$$

at a suitable grid of points, but this approach seems somewhat artificial with inherent extrapolation problems that might provide deceptive results.

To form d, one treats $(Q_1/(n+1), R_1/(n+1)), \ldots, (Q_n/(n+1), R_n/(n+1))$ as a random sample from a bivariate uniform distribution and then

applies one of the density estimation techniques of Chapter 3 that permit bivariate generalizations. Two techniques will be mentioned here for obtaining $H(d)$.

One approach is to subjectively decide upon the best window width or smoothing parameter to produce a kernel or nearest neighbor bivariate density estimate $\tilde{d}(u_1,u_2)$. One observes that such a subjective approach unduly complicates the procedure, but such problems cannot be overcome. The second approach is to form $\hat{d}(u_1,u_2)$ using $\tilde{d}(u_1,u_2)$ as a dependent variable in the regression approach of the last section. One then may form estimator (4.5.5) or (4.5.6). The parametric representation afforded by the regression approach makes method two less computationally cumbersome than it would be in the first approach. Hence, one uses (4.5.5) for a suitable choice of d and (4.5.6) for the parametrically smoothed bi-information density.

A Monte Carlo study of $H(\tilde{d})$ and $H(\hat{d})$ has been carried out for 100 iterations of samples of size 50 and 100 for a bivariate normal distribution and a distribution composed of one standard normal marginal and another marginal corresponding to a conditional Cauchy(0,1) distribution. The sensitivity of Pearson's r to outliers is well documented and hence is not investigated here. The nonparametric procedures have well known robustness properties which are mimicked by the entropy statistics since the latter are constructed using ranks of a "trimmed" data set. (A discussion of the computer algorithm generating $H(\tilde{d})$ and $H(\hat{d})$ may be found in the next chapter.) Table 3 presents several quantile values for the various entropy statistics for the sample sizes 50 and 100 obtained from the

simulations. The notation H(d8) refers to an order 8 expansion, meaning that all 8 bivariate combinations of the indices (-1,0,1) were included. H(d24) contains all 24 combinations of (-2,-1,0,1,2), etc. Note that in each case the index (0,0) is excluded since the constant

Table 3. Quantiles for Entropy Statistics

n	p	$H(\tilde{d})$	H(d8)	H(d24)	H(d48)	$H(d_0)-H(d)$
50	0.01	-0.222	-0.229	-0.293	-6.289	-0.206
	0.05	-0.137	-0.179	-0.226	-5.294	-0.163
	0.10	-0.100	-0.169	-0.202	-4.477	-0.144
	0.25	-0.067	-0.145	-0.179	-2.838	-0.102
100	0.01	-0.251	-0.169	-0.197	-0.248	-0.029
	0.05	-0.181	-0.139	-0.179	-0.204	-0.002
	0.10	-0.157	-0.128	-0.163	-0.197	0.018
	0.25	-0.132	-0.107	-0.149	-0.174	0.048

term has been incorporated into the integrating factor.

Power studies were conducted for various values of ρ for the normal sample and for a sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ generated by Y=X+C where X is a standard normal random variable and C is a Cauchy(0,1) random variable. This model corresponds to a general regression model with Cauchy errors. It may be shown that Y is positively regression dependent on X since the conditional distribution of Y given X=x is a Cauchy with median x. Hence, only the normal theory statistic r will have assumptions violated for this case.

The results for the power study may be found in Table 4. For n=50, the entropy statistics are disappointing in comparison with the

Table 4. Monte Carlo Results for Power Study of Measures of Association

significance level=0.10

N	RHO	r	ρ_n	τ_n	$H(\tilde{d})$	H(d8)	H(d24)	H(d48)
50	0.2	0.49	0.44	0.46	0.31	0.11	0.22	0.16
	0.4	0.90	0.91	0.91	0.58	0.21	0.23	0.09
	0.6	1.00	1.00	0.99	0.71	0.31	0.32	0.15
100	0.2	0.67	0.66	0.65	0.37	0.18	0.17	0.19
	0.4	0.98	0.98	0.98	0.53	0.51	0.48	0.42
	0.6	1.00	1.00	1.00	0.98	0.91	0.90	0.86
Cauchy 100		0.32	1.00	1.00	0.90	0.95	0.96	0.94

For Cauchy sample, the g.o.f. statistic H(d-Normal)-H(d) has power 0.88.

correlation statistics for local alternatives to $\rho=0$ in the normal case, although $H(\tilde{d})$ is fairly competitive. This suggests that the density estimation approach should not be recommended for small samples, particularly using the numerical integration statistics. One suspects that the problem of extrapolation has unduly weakened the effect of the statistics, while for $H(\tilde{d})$ no extrapolation is attempted. For n=100, the results are more promising, and for the non-normal case, the entropy statistics perform well compared to Spearman's Rho and Kendall's Tau, and, as expected, greatly surpass the normal theory statistic r. For this study, we also included the goodness-of-fit statistic

$$H(d_0) - H(\tilde{d}) = -(1/2)\log(1-r^2) - H(\tilde{d}) \qquad (4.5.7)$$

whose quantile values obtained for the normal cases are also given in Table 3. The power result of .88 suggests that this entropy based statistic may be competitive with existing procedures for testing bivariate normality. However, further simulations are warranted.

The poor performance of the entropy statistics suggests that some modification be employed to overcome the consistency and power problems. Rather than employ a numerical Riemann integral, an alternate approach is to consider the numerical Lebesgue integral for orders 8, 24, and 48. Recall, the raw entropy statistics based on the nearest neighbor estimate is a Lebesgue integral w.r.t. the empirical c.d.f. To obtain a numerical approximation of the Lebesgue integral for the minimum information estimators, the estimated dependence density is evaluated at a k by k grid of points in the unit square and then is treated as a vector of dimension k^2. One then forms the corresponding vector of d log(d) values and obtains a robust measure of location such as the trimean that serves as a numerical approximation to the Lebesgue integral. The amount of calculations involved prohibit a large scale simulation of this approach, but limited experience with some of the data sets considerd in Chapter 6 are promising at least for cases involving small correlations. The corresponding use of quantile techniques to analyze the vector of d and d log(d) values may aid in determining an appropriate order of expansion.

Without supporting theory, a general simulation study for a wide

class of alternatives to bivariate normality is infeasible. The results of Vasicek (1976) are promising but need to be extended to the bivariate case. For the simulations performed, the nearest neighbor density d was computed using $k(n)=5$. In practice, one might try a variety of values of $k(n)$ to arrive at a "pleasing" shape, and then examine the entropy measures. Unfortunately, this subjective approach cannot be incorporated into a simulation study.

4.6 Other Applications

4.6.1 Nonparametric Regression

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample from a bivariate distribution with c.d.f. $F_{X,Y}$, p.d.f. $f_{X,Y}$, and associated marginal and conditional functions with the usual notation. One often attempts to discern a relationship between X and Y in order to predict Y given a value of X. An important object in this cases if the <u>regression function</u>

$$r(x) = E[Y|X=x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy. \qquad (4.6.1)$$

From the definition of the conditional p.d.f. one may express (4.6.1) by

$$r(x) = h(x)/f_X(x) \qquad (4.6.2)$$

where

$$h(x) = \int_{-\infty}^{\infty} y\, f_{X,Y}(x,y)\, dy. \qquad (4.6.3)$$

Watson (1964) and Nadaraya (1964) as referenced in Cheng and Taylor (1980) used representation (4.6.2) and kernel density estimation results to suggest estimating $r(x)$ by

$$\hat{r}(x) = \hat{h}(x)/\hat{f}(x) \qquad (4.6.4)$$

where $\hat{f}(x)$ is the kernel density estimate of $f(x)$ given by

$$\hat{f}(x) = \int_{-\infty}^{\infty} [1/h(n)]\, K[(x-x')/h(n)]\, dF_n(x')$$

$$= [1/nh(n)] \sum_{j=1}^{n} K[(x-X_j)/h(n)] \qquad (4.6.5)$$

and $\hat{h}(x)$ is related to the kernel estimate by

$$\hat{h}(x) = \int_{-\infty}^{\infty} y\, [1/h(n)]\, K[(x-x')/h(n)]\, dF_n(x',y)$$

$$= [1/nh(n)] \sum_{j=1}^{n} Y\, K[(x-X_j)/h(n)]. \qquad (4.6.6)$$

Rosenblatt (1971) gives properties of $r(x)$ and Cheng and Taylor (1980) extend these to a more general case of a k-dimensional X-vector. This technique is completely general and can be thought of as taking a weighted average of Y values based on the X observations. This is

easily seen by expressing (4.6.4) as

$$\hat{r}(x) = \sum_{j=1}^{n} K[(x-X_j)/h(n)]Y \; / \sum_{j=1}^{n} K[(x-X_j)/h(n)]. \qquad (4.6.7)$$

The kernel K acts as a focusing function giving more weight to Y values for X_j in a neighborhood of x.

The representation (4.6.4) suggests a multitude of estimators based on the various nonparametric density estimation techniques of Chapter 3. Asymptotic properties may be intractable for many of these cases, however. Nonetheless, one may seek to rewrite (4.6.4) to permit application of some of the quantile based techniques mentioned previously.

In Chapter 2 we observed that (4.6.1) could be translated to a regression quantile function by the formula

$$rQ_X(u_1) = \int_0^1 Q_Y(u_2)\, d(u_1,u_2)\, du_2, \qquad (4.6.8)$$

where $d(u_1,u_2)$ is the dependence density. This formula is derived from (4.6.2) with the transformation $x=Q_X(u_1)$ and $y=Q_Y(u_2)$. One obtains

$$rQ_X(u_1) = \int_0^1 Q_Y(u_2)\, [f_{X,Y}(Q_X(u_1),Q_Y(u_2))/f_X(Q_X(u_1))]\, dQ_Y(u_2)$$

$$= \int_0^1 Q_Y(u_2)\, [f_{X,Y}(Q_X(u_1),Q(u_2))/f_X(Q_X(u_1))]\, q_Y(u_2)\, du_2 \qquad (4.6.9)$$

which then simplfies to (4.6.8) by virtue of the reciprocal identity.

Estimation of (4.6.8) may be obtained by numerically integrating products of sample quantile and dependence density functions. For example, selecting a grid $u_{21},\ldots,u_{2m}$ of equally spaced values between 0 and 1, one obtains the Riemann sum

$$\hat{rQ}_X(u_1) = (1/m) \sum_{j=1}^{m} \hat{Q}_Y(u_{2j})\ \hat{d}(u_1,u_{2j}) \qquad (4.6.10)$$

as an estimator of the regression function. A regression curve may then be plotted for various values of u_1.

Parzen (1979a) emphasizes the conditional quantile function in approaches to nonparametric regression. He considers versions of (4.6.7) in the quantile domain with emphasis on smoothing raw regression function estimates based on empirical quantile functions. In Parzen (1977), raw estimates of the partial derivative of $D(u_1,u_2)$ w.r.t. u_1 provide alternatives to (4.6.10) for estimating $rQ_X(u_1)$, but this is proposed only as a "quick and dirty" technique. Our emphasis on obtaining smooth estimates of $d(u_1,u_2)$ should make (4.6.10) the preferred estimator of the regression quantile function, but in any case, asymptotic properties remain to be investigated.

We have only examined nonparametric regression from a density estimation approach. Huber (1981) suggests robust least squares procedures, and Hajek and Sidak (1967) consider some linear rank tests for hypothesis testing concerning linear regression coefficients. We have avoided any assumption of linearity in our discussion, but when such an assumption is justified, nonparametric approaches to the linear regression problem may be preferred. In a pure modeling

approach, one would desire some sort of residual analysis that tests for white noise of residuals in an effort to evaluate the model given by (4.6.10).

4.6.2 Discrimination and Classification

Often one seeks to classify an individual with bivariate characteristics (X,Y) into one of two parent populations. If data is available on each population, one seeks to classify (X,Y) into the population that seems most closely related to (X,Y). If the bivariate p.d.f.'s $f_1(x,y)$ and $f_2(x,y)$ of populations 1 and 2, respectively, are known, one forms the likelihood ratio

$$\lambda(X,Y) = f_1(X,Y)/f_2(X,Y) \qquad (4.6.11)$$

and classifies (X,Y) into population 1 if $\lambda(X,Y) \geq 1$ and into population 2 otherwise. If the two populations have normal distributions with common unknown covariance matrix Σ and different unknown mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$, one obtains the corresponding sample estimates of these quantities and forms the sample discriminant function

$$W = [\underline{X} - (1/2)(\underline{\bar{X}}_1 + \underline{\bar{X}}_2)]' \, S^{-1} \, [\underline{\bar{X}}_1 - \underline{\bar{X}}_2] \qquad (4.6.12)$$

where $\underline{X}' = (X,Y)$, $\underline{\bar{X}}_1' = (\bar{X}_1, \bar{Y}_1)$, etc. One then assigns (X,Y) to population 1 if W>0 and to population 2 otherwise. Morrison (1976) gives an adequate description of this normal theoretic approach.

If the assumption of bivariate normality cannot be justified, one may seek to estimate the unknown densities of each population to form

$$\hat{\lambda} = \hat{f}_1(X,Y)/\hat{f}_2(X,Y) \qquad (4.6.13)$$

and proceed as before. An alternate approach is suggested working in the quantile domain. For a sample from population 1, one estimates $d_1(u_1,u_2)$ and univariate c.d.f.'s F_X and F_Y. For population 2, estimates of $d_2(u_1,u_2)$, G_X, and G_Y are obtained. One then forms

$$\hat{\lambda} = \hat{d}_1(\hat{F}_X(X),\hat{F}_Y(Y))/\hat{d}_2(\hat{G}_X(X),\hat{G}_Y(Y)) \qquad (4.6.14)$$

and proceeds accordingly. Observe, if in sample 1, $X_{(k)} \leq X \leq X_{(k+1)}$, $\hat{F}_X(X) = k/(n+1)$ is an acceptable raw estimate for $U_1 = F_X(X)$. One may prefer using estimates of the bivariate density-quantile function over using estimates of the dependence density. Such an approach, however, seems to assume equal marginals for both populations. While probabilities of misclassification based on (4.6.14) may seem difficult to obtain, this approach "exhausts the data" by utilizing all of the relevant sample functions (if only indirectly) in creating the discriminant function. Thus, the approach would seem more sensitive than an approach dealing only with likelihood ratios in the density domain, and hence one might expect small probabilities of misclassification using this technique. This remains an open research question.

4.6.3 Parametric Modeling

A useful extension of the minimum information density estimation technique concerns estimating parameters of a parametric model for the purpose of ascertaining the adequacy of the model. Consider the canonical exponential model of order m given by

$$\log f(x) = \sum_{k=1}^{m} \theta_k T_k(x) - C(\theta_1,\ldots,\theta_m), \qquad (4.6.15)$$

where $T_1(x),\ldots,T_m(x)$ are called sufficient statistics. This density maximizes entropy subject to the constraints

$$\int_{-\infty}^{\infty} T_k(x)\ f(x)\ dx = \tau_k,\ k=1,\ldots,m, \qquad (4.6.16)$$

where $\tau_1,\ldots,\tau_k$ are called moment parameters. This implies that the normal distribution maximizes entropy over all other distributions with specified mean and variance. Using the minimum information approach for the parametric model (4.6.15), one obtains least squares estimates for the model parameters which can lead to estimates of the moment parameters. Recall, from the theory of exponential models, one has

$$(\partial/\partial\theta_k)\ C(\theta_1,\ldots,\theta_m) = \tau_k, \qquad (4.6.17)$$

and hence under suitable regularity conditions the moment estimators

$$\hat{\tau}_k = (1/n) \sum_{j=1}^{n} T_k(X_j), \quad k=1,\ldots,m, \tag{4.6.18}$$

are also maximum likelihood estimators and thus the estimators

$$\hat{\theta}_k = \hat{\theta}_k(\hat{\tau}_1,\ldots,\hat{\tau}_m), \quad k=1,\ldots,m, \tag{4.6.19}$$

are m.l.e.'s by the invariance property. Hence, one may form the least squares estimates and compare these to the m.l.e.'s for a diagnostic check of the adequacy of the parametric model given by (4.6.15).

Example 4.6.1 For the normal case, $T_1(x)=x$ and $T_2(x)=x^2$, so that $\tau_1=\mu$ and $\tau_2=\mu^2+\sigma^2$. The canonical parameters are given by $\theta_1=(\mu/\sigma^2)$ and $\theta_2=-1/(2\sigma^2)$. The regression approach would form estimators of the parameters in the model

$$\log f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \tag{4.6.20}$$

with a stochastic element introduced when the nearest neighbor estimate replaces $f(x)$ in the model.

Example 4.6.2 For a gamma model with density

$$f(x) = [b^a/\Gamma(a)]\, x^{a-1}\exp(-bx), \tag{4.6.21}$$

where $x>0$, $a,b>0$, the canonical form is

$$\log f(x) = \theta_0 + \theta_1 \log x + \theta_2 x, \qquad (4.6.22)$$

where $T_1(x)=\log x$ and $T_2(x)=x$ with $\theta_1=(a-1)$ and $\theta_2=-b$. For a location-scale gamma model, problems are encountered unless the location parameter is known. For unknown location, one estimates it by the minimum value of the data and then treats this as the true parameter value to be able to obtain the least squares estimates for a and b.

This parametric modeling approach will be illustrated in Chapter 6 applied to normal and gamma data. Note that this technique is not recommended for parameter estimation when the model is known to be valid, but instead is suggested as a method for checking the adequacy of a model. One may wish to investigate distributional properties to suggest inferential goodness-of-fit procedures. Bivariate extensions are fairly straightforward and will not be considered here.

4.7 Concluding Remarks

Perhaps the greatest weakness of nonparametric statistics until now has been its failure to adequately handle multivariate problems. The problem seems to center around the insistence upon carrying out inferential procedures on parameters of a probability model and the inability to nonparametrically estimate these parameters. For example, the contrasts of interest in an analysis of variance setting often rely on robustness properties in the absence of nonparametric

multiple comparison procedures. Heuristic solutions (such as replacing the data by ranks) may appear to work in some cases, but further study is warranted.

We have followed the recent approach of placing function approximation ahead of parameter estimation. Naturally this is a more difficult estimation problem, but once solved, it leads directly to a solution to estimating parameters of interest. Unfortunately, most techniques must reside in the category of the exploratory rather than confirmatory because of the lack of theory to support the procedure. The acceptance of Monte Carlo studies has gradually improved over the years, but unfortunately, function approximation does not lend itself to general Monte Carlo experimental designs especially when seeking comparisons for a wide class of alternatives. Many of the expansion techniques have adequate theoretical motivation, the main problem being that of order determination. Consequently, techniques such as the autoregressive approach of Parzen or the orthogonal expansion technique of Kronmal and Tartar that exhibit objective order determining criterion are the most promising of those suggested in the literature. This motivated the expansion techniques considered in this chapter. Unfortunately, the criteria employed did not seem to perform as well as hoped, but once a suitable criterion is obtained, the minimum information technique will become even more useful than approaches such as the kernel method that necessitate examining multitudes of shapes to arrive at a conclusion. Furthermore, the generalization of methods based on the dependence density to a multivariate setting are fairly straightforward especially in the

nonparametric regression framework. The sample size required in a multivariate setting to adequately perform function approximation will always pose a serious problem, however, and hence parametric approaches will continue to dominate small sample settings when one can justify the assumed model to any degree of satisfaction.

5. COMPUTER SOFTWARE FOR BIVARIATE DATA ANALYSIS

5.1 Introduction

The major statistical computer packages have yet to fully enter the fields of nonparametric density estimation or bivariate data analysis. Consequently, one must create his own computer programs to carry out many of the procedures detailed in this work. For nonparametric density estimation, most statistical packages will have histogram procedures, but only IMSL (International Mathematical and Statistical Libraries) provides routines to do other types of nonparametric density estimation. When one ventures too far from classical normal theory procedures or some of the more popular nonparametric techniques, the existing statistical software packages are of little help.

Ideally, the examination of density curves is carried out in an interactive computing environment so that shapes can be examined and adjusted quickly to arrive at an "optimal" choice for the estimated density. However, the programs we will discuss were written in FORTRAN for batch processing. This was done for a variety of reasons which will not be discussed here. The translation of FORTRAN code into an interactive language such as BASIC is not too difficult, and some systems have time sharing FORTRAN capabilities. The system employed for our procedures has a "simulated" interactive language that permits quick access to batch output at a CRT terminal. The computing environment for program implementation will be discussed

later.

The writing of computer programs is considered to be an art by some, and a particular program often mirrors something of the personality of its creator. Thus, examination of the code that we have written illustrates a certain philosophy of programming that will be discussed in the next section. The actual programs that we have written will be discussed in sections 5.3 and 5.4. We will conclude this chapter by examining the facilities that were used and the typical effort required to execute a program and retrieve the results.

5.2 A Philosophy of Statistical Computing

There are many ways to attack the writing of computer code to carry out some desired purpose. The recent popularity and utility of structured programming has caused it to be a widely practiced form of program construction. The idea behind this approach is to carefully organize a program so that it flows smoothly from one computation to the next without haphazard placement of loops and branches. There are a variety of ways to organize a program with this approach in mind. One method is to create a bank of subroutines each of which is carefully designed to carry out a specific task, and then write a fairly terse main program that systematically accesses these routines. Using this approach, one may discover that efficient routines already exist that perform certain tasks, and hence one need not expend effort in creating the routine oneself. The IMSL FORTRAN subroutine library contains many useful techniques backed by extensive testing that could

probably not be matched by a programmer with limited resources. Many systems maintain a variety of subroutine libraries and catalogued procedures that may be useful to programmers. With a collection of tested subroutines at ones disposal, the trauma of debugging a large program is greatly reduced. With this in mind, one may wish to insert checks and flags in a routine to guard against its misuse in later applications. An alternate philosophy adopted by some is to write completely self contained main programs that systematically perform every task in the main body of the program. Careful documentation of such programs make it easy for one to examine the code to discover what tasks the program is performing. Arguments do not have to be passed back and forth to subroutines and array dimensioning is handled only once without the need to trace dimension values throughout a program. While this approach has some advantages, the major weakness is that a great deal of repitition may occur in writing the program so that less effort may be spent thinking of new approaches or ways of making the program more efficient. We favor the subroutine approach as one will soon realize upon examining our programs.

There are three goals inherant in computer program construction.

1) The program should correctly perform the task for which it was intended.
2) The program should work, i.e., it should be reliable, anticipating any awkward contingencies that might occur.
3) The program should be easy to read and easy to maintain.

A fourth goal is often added to this list.

4) The program should be portable, i.e., it should be designed to work in a general computing environment for a variety of computer systems.

Using a popular language such as FORTRAN or COBOL and avoiding machine dependent conventions should promote portability, although most likely some translation will be needed when going from one system to another. Since we anticipate that our programs may be used at more than one computer installation, we have made some attempt to avoid machine dependent conventions.

The goals presented above are important, but as with the theory vs. simulation dilemma, one can never anticipate the infinite possibility of data sets that may be exposed to a program. Therefore, anticipating such problems as division by zero may promote the efficient use of a program, especially if an error might occur in a minor step not crucial to the general task of the program. Insertion of options that default to logical values will also help insure that a program completes its task despite minor errors of no consequence.

A wide variety of computer languages exist each aimed at empasizing a particular application. FORTRAN is designed for scientific computations while COBOL is geared more towards business applications. Thus COBOL may be better suited for character manipulation while FORTRAN might be preferred for "number crunching". Since our main goal is one of computation, we have chosen to write our

programs in FORTRAN. Also, as suggested, a FORTRAN program may written to be fairly portable. The only problems might occur in certain format conventions or specialized functions such as array manipulation. Furthermore, the popularity of FORTRAN will make the code understandable to most users in case they wish to make modifications tailored to their own specialized applications. Although many options are included in our programs, a user may wish to add or delete options to reflect the specialized environment in which he is working. The FORTRAN language has many qualities to recommend it which need not be discussed here.

Finally, we note that while efficiency is always an important concern especially in the construction of a large program, one may be more concerned with documenting and organizing a program so that it may easily be used by others. Naturally, one seeks a suitable compromise between efficiency and ease of use so that a program is not prohibitively expensive or time consuming. Our first concern is for accuracy and precision. When these attributes are sufficiently safeguarded, then one may search for ways to make a program more efficient. Clearly one does not desire a program that quickly and efficiently computes the wrong answer, although this is a common occurence in computer applications. More insight may be gained into this philosophy by a closer inspection of the routines we have written. Some comments will be made about obstacles that had to be overcome, and references will be made to the authors of contributing routines. For a discussion of statistical computing one may consult Kennedy and Gentle (1980), and for a general discussion of the science

and art of computing, Knuth (1968) is a useful reference.

5.3 Univariate Density Estimation Routines

The popularity of the histogram makes it readily available from many statistical computer packages. The Statistical Analysis System (SAS, 1979) provides histograms in its CHART procedure. The BMDP Biomedical Computer Programs (BMDP, 1979) provide two programs, BMDP2D and BMDP5D, that produce histograms for a data set. MINITAB (Ryan, Joiner, and Ryan, 1975) has a command HISTOGRAM that will produce a histogram for a specified data vector. The Statistical Package for the Social Sciences (SPSS, see Nie, et al., 1975) provides a histogram through the procedure FREQUENCIES. All of these routine have an objective default for computing cell widths and boundaries. For example, PROC CHART of SAS by default will let $m=FLOOR[1+3.3\ log(n)]$ where FLOOR is the greatest integer function and n is the sample size. The range of the data is then divided into m equally spaced intervals yielding h=range/m.

The histogram is the only form of density estimation available from most packages. IMSL has two routines that offer alternatives to the histogram, but it is the only major source of such routines. Kernel estimation is performed by the IMSL routine NDKER for user provided kernel and specified window width. The routine NDMPLE performs discrete maximum penalized likelihood density estimation for user specified smoothing parameter. These routines are well documented and easy to use.

For alternatives to the existing software, one may consult the literature to obtain algorithms to be programmed. The moment techniques of Cencov (1962) are easily programmed once one has surmounted the problem of generating orthogonal systems of functions in L^2 space. Tartar and Kronmal (1976) provide a very readable account of implementing density estimation techniques. Based on these references and the use of such numerical algorithm sources as Abramowitz and Stegun (1972), one may readily construct FORTRAN routines to perform density estimation.

For the minimum information techniques developed in Chapter 4, one may use existing regression software to implement the procedures. However, for complex exponential systems some adjustments may have to be made. The TIMESBOARD FORTRAN library of Newton (1979) contains some useful routines for handling complex regression in addition to many general purpose routines. An alternative to obtaining FORTRAN regression routines to implement this procedure is to create a "regression data set" and use this as input to a procedure such as PROC GLM of SAS. Other regression software may also be employed.

The author has written five FORTRAN routines to perform univariate nonparametric density estimation. These routines along with the IMSL routine NDKER will be applied to several data sets in the next chapter to illustrate their use. The five routines we have written are called NNDEN (Nearest Neighbor), KTDEN (Kronmal-Tartar type with trigonometric polynomials), TKDEN (Tartar-Kronmal type with complex exponentials), MIDEN (Minimum Information type with Legendre polynomials), and CMPDEN (minimum information type with CoMPlex

exponentials). The routines NNDEN, KTDEN, and TKDEN are easily written following the algorithms described in the literature and hence will not be listed here. The routines MIDEN and CMPDEN are listed in Appendix A. Appendices C and D contain a collection of subprograms accessed by these procedures.

The density estimation routines were written to accept standard input and produce standard output within the routine, passing as few arguments back and forth as possible. Our purpose for writing the routines was to get plotted output quickly and efficiently for a variety of smoothing parameter values. For more practical applications, one may wish to pass the actual estimated density values back to the main program to be used for further investigation or analysis. This is easily accomplished by modifying the calling arguments. The only problem might be in controlling the values at which the density is evaluated, but routines are available for interpolation if necessary. For the parametric orthogonal expansion models, one need only pass paramter estimates with the corresponding variable indices back to the calling program to be employed as needed. As suggested, this is an advantage in using expansion techniques, namely, that one need only knowledge of a few paramter estimates to completely describe the estimated density rather than knowledge of function estimates for a large number of values. For example, a vector of size 10 may store all of the relevant information about an orthogonal expansion approximation, while vectors of size 50 or more are usually required for nearest neighbor, kernel, or DMPL type estimates, depending upong the number of estimated values one wishes.

Typically, the unique feature of each density estimation routine is the algorithm employed to derive the estimate. Otherwise, each routine has a general framework. This framework is outlined as follows:

Input: Data (X), sample size (N), minimum value of data (A or XMIN), maximum value of data (B or XMAX), and options (IOPTk).

Preprocess Data: If data is modified by an algorithm, let Y(I) = X(I) and use Y vector in the procedure. If trimming or scaling is required, perform necessary transformations before exposing data to the algorithm. (Usually standardizing to an interval (a,b) is performed within the algorithm to reduce the lines of computer code required.)

Invoke Algorithm: Expose the (possibly transformed) data to the algorithm to obtain parameter estimates or estimated density values. The standard approach is to then obtain density estimates evaluated at N or 100 equally spaced values between XMIN and XMAX.

Compute Density Functionals: Two versions of each routine exist, one applied to data with unknown distriburion and one applied to data with a specified null distribution. For data with unknown distributions, estimates of the mean, variance, and mode are obtained by numerical integration and grid search techniques

performed on the estimated dens'ty. For a specified null distribution, estimates of integrated squared error, mean squared error, and maximum absolute deviation are obtained in a similar fashion. The options allowed for null distributions include normal, gamma, and a mixture of two normal densities since these are the cases considered in Chapter 6.

Plot Density Estimates: For unknown parent distributions, a printer plot is obtained displaying equally spaced X values, estimated density values, and the corresponding shape of the estimated density. For specified null distributions, an overlay plot is obtained with an additional listing of the null density values. Examples of these plots will appear in Chapter 6.

The options IOPTk usually involve choices of smoothing parameter orders or null density values. They may also decide the number of different estimates to be obtained. Output of plots and parameter estimates is automatic, but the procedures may be modified by the user to control output.

For the NNDEN procedure, if one wishes to obtain density estimates at m points, roughly 4mn+kn+6 computations will be performed for a data set of size n and smoothing paramater k=k(n). The most involved step is finding the smallest radius r such that a sphere with radius r centered at the evaluated data point contains k-1 additional points. To perform this computation, (n-1) radii are computed and k calls to routine MIN are made with Y(IMIN) being replaced by a large

positive value each time. It is conceivable but unlikely that all radii will be larger than this value, which is currently set at 7.0E+75. This is near the machine limit for single precision constants and represents 7 followed by 75 zeros. Hence, it should be suitable for most data sets. In the univariate case, increasing the value of k does not severely lengthen the procedure, but the multivariate analog can be unduly lengthened by large choices of k due to the involved computations for the volume of a hypersphere.

The algorithms employed for the KTDEN and TKDEN procedures are perhaps the easiest to program. The most difficult step is generating a system of orthogonal functions. Since KTDEN and TKDEN compute moment estimators based on orthogonal systems of with pleasing forms, the algorithms for generating density estimates are easily programmed. The trigonometric polynomials or complex exponentials may be programmed directly into the averaging routine without problems of generating coefficients as is the case if Legendre [-1,1] polynomials had been used. The data must be standardized to (0,1), but this is easily implemented directly into the algorithms using the transformation (Y(I)-A)/(B-A) where A=min(X) and B=max(X) and Y is the vector of X values described above. Estimates are then computed using a truncated order MN=m(n), with the maximum order currently set at 10. The vector THETA of moment coeficients computed for the orthogonal expansion need only be computed once up to the maximum order by virtue of the orthogonality property. Estimates of varying orders may then be obtained by simply calling for the coefficients needed. The routine TKDEN computes a best order using the MISE criterion, while

KTDEN merely displays plots for user specified orders.

The routine KRDEN merely generates the input parameters for the IMSL routine NDKER which is documented in the IMSL Library, Volume 2 (1980). Various values of H=h(n) may be employed to obtain different shapes for the estimated density. The kernel employed is the normal kernel programmed in the function subprogram XNKER.

The routines MIDEN and CMPDEN are the most complicated routines in that a variety of regression subprograms must be employed to generate the parameters of interest. MIDEN uses a subprogram called LEGP to generate the matrix COF of Legendre polynomial coefficients. CMPDEN has the advantage shared by KTDEN and TKDEN of employing the orthogonal functions directly in the algorithm. The difference between these techniques and the above orthogonal expansion techniques is that a covariance matrix COV (PHI in CMPDEN) must be computed and supplied to a least squares algorithm to obtain least squares parameter estimates for the truncated orthogonal expansion. A SWEEP operator is employed in the sequential regression routine SEQREG to obtain the coefficients in MIDEN, while a complex SWEEP operator is used in CSQREG for CMPDEN. The SWEEP operator has many computational advantages. Kennedy and Gentle (1980) discuss some of its properties.

Both MIDEN and CMPDEN use the logarithm of an initial density estimate from NNDEN to serve as a dependent variable in the regression framework. A value of 8 is used to m(n) as a default, but was modified for some of the runs described in Chapter 6. For comparison purposes, a version of CMPDEN was written that used NDKER in place of NNDEN to obtain initial estimates, and while results were excellent

for the special case considered, some reasons will be given in Chapter 6 why this practice is not recommended. Note that the maximum order of expansion is set at 10 for both procedures and that orthogonality with respect to the initial estimate has not been induced. This means that addition of terms in the expansion will change all of the previously computed coefficients. Consequently, one may wish to modify this procedure accordingly, but one will note in the next chapter that non-orthogonality does not pose a serious problem.

In programming the complex regression technique for CMPDEN, several early failures emphasize the importance of taking care when dealing with complex values. Although the final density estimate will be real if one employs conjugate pairs in the expansion, the erroneous deletion of the imaginary part of the estimated coefficients invalidates the obtained estimate. One should avoid the suppression of imaginary terms to obtain real valued estimates until the procedure has been thoroughly tested. Examinations of the results of Chapter 6 reveal that the technique for obtaining estimates in both TKDEN and CMPDEN produces complex valued coefficients. When applied properly in conjugate pairs, the imaginary terms always vanish.

The problem of order determination has been considered in both TKDEN and CMPDEN. As mentioned, the MISE criterion is used to obtain a "best" order for the expansion in TKDEN. The AIC criterion is computed in CMPDEN to diagnose a best order, but in practice this criterion does not seem to perform as well as the MISE criterion.

A parametric version of MIDEN called NPDEN was created specifically to handle least squares estimation of the parameters in

the canonical exponential model representation of gamma and normal densities. The independent variables are log x, x, and x^2 in the expansion and the coefficients are computed by inserting the appropriate variables into the regression model.

The utility of the minimum information regression approach is illustrated by the two step FORTRAN-SAS program listed in Appendix A.3. A FORTRAN PROGRAM uses LEGP to generate independent variables and NNDEN to generate the dependent variable to be used in the SAS procedure GLM. Predicted values are written into a data set called TWO, and in a DATA step the exponent of the predicted values is obtained and submitted to the procedure PLOT where the estimated density is then plotted against the original X value. This two step procedure exemplifies the ease in adapting the minimum information regression approach to existing regression software. More difficulty may be encountered in the use of complex exponentials because complex regression routines are not to be found in the major statistical packages.

Appendix C contains some useful subprograms accessed by the above routines. Appendix D contains several of the plotting subprograms employed to obtain printer plots of the estimated densities. PLOTXY generates 45 equally spaced X values and generates the corresponding Y values by linear interpolation. Then a plot of Y as a function of X is produced with the values of X and Y printed for each plotted point. Corresponding to this routine is PLTXYZ which produces an overlay plot of Y and Z as functions of X. The routine PPLOT produces a more appealing plot but presents only a scale of X and Y values. PPLOT is

recommended for scatter plots as it does not induce linearity by interpolating values as does PLOTXY. For more elaborate plotting using the Versatec plotting software, the subprogram GPLOT of the TIMESBOARD Library (Newton, 1979) is a flexible multiple purpose routine for obtaining high resolution plots. SAS/GRAPH (1981) also contains many useful plotting procedures. A version of NPDEN accesses GPLOT to obtain plots on the Versatec Electrostatic Plotter. Some of the three dimensional plotting procedures of SAS/GRAPH will be mentioned in the next section.

5.4 The BISAM Bivariate Data Modeling Program

The program BISAM is a general purpose bivariate data modeling program designed to provide a variety of univariate and bivariate descriptive statistics and graphical output. The core of BISAM is the routine CMPINF (CoMPlex regression using INFormation functionals) which serves as a generalization to CMPDEN for the bivariate case. Within CMPINF the rank transformed data is exposed to a bivariate version of NNDEN with computations proceeding as in CMPDEN. A modular arithmetic scheme is devised to imbed the two-dimensional subscripting of coefficients into a one dimensional indexing scheme. A maximum order of 7 is specified and a 49 by 49 covariance matrix (omitting the constant term) is then supplied to CSQREG described above. Currently, orders 8, 24, and 48 are automatically supplied with the user having the option to override these values with user-supplied even orders (to insure that the estimated dependence density is real). A listing of

the main body of BISAM along with the major subprograms employed may be found in Appendix B. The minor and peripheral subprograms employed may be found in appendices C and D.

The main calling program of BISAM reads the bivariate data set with the variables read in one at a time as univariate data sets. The first card of a data set should contain a descriptive title describing the data set. The second card should contain the number of data points N and the format of the input variables. The format that reads the second card is given by (I5,4X,5A4). Coded in the appropriate format, the N data points should then follow. When BISAM finishes reading the two data sets, it immediately checks to see if the two sample sizes agree. If they do not, the program terminates. Before reading the two data sets, however, the user must specify several options to be employed in the analysis. Hence, the first data card will be an option card with the following input options in (9I5) format:

NTAPE - tape where data set resides

MORD - maximum autoregressive order to be used for univariate density estimation (≤6)

IDQX,IDQY - null distributions for autoregressive smoothing:

1 = Normal	4 = Double Exponential
2 = Exponential	5 = Uniform reciprocal
3 = Logistic	6 = Cauchy

A value of 1 is recommended. For a more complete listing and description, see Parzen and Anderson (1980).

IPLT1 - Scatter plot options:

0 = no scatter plots

1 = scatter plot of data

2 = scatter plot of rank transformed data

3 = both scatter plots

IPLT2 - Univariate density plotting options:

0 = no quantile box plots

1 = produce quantile box plots

IDST - Univariate descriptive statistics:

0 = no descriptive statistics displayed

1 = descriptive statistics computed and displayed for each variable

KDEL - Maximum number of extreme points to exclude from analysis.

Extreme points are determined by distance from the median, and if X and Y extreme points correspond, they count as two points although only one will be excluded from the analysis. Hence, KDEL usually exceeds the actual number of points so deleted.

These options illustrate the variety of output available from BISAM. The univariate descriptive statistics, plots, and autoregressive density estimates are obtained by employing some modified routines from ONESAM (Parzen and Anderson, 1980). The routine PPLOT described in the last section is used to obtain the scatter plots, while PLOTXY is used to plot the univariate density-quantile functions.

Before calling CMPINF, several routines are accessed that provide some of the standard correlation statistics discussed in section 6.3. CMPINF then uses the various estimates of the dependence density to obtain entropy measures of association. These are displayed along with the standard correlation statistics at the end of the output. Intrmediate output consists of univariate descriptive statistics and plots (if requested) and coefficients for the expansion of the logarithm of the dependence density. An integrating factor is also displayed providing a diagnostic as to the legitimacy of the estimated dependence density. Sample output from BISAM will be presented in the next chapter.

The current version of BISAM writes the values of an estimated dependence density and bivariate density-quantile function to a temporary disc file to be accessed by a SAS/GRAPH procedure. PROC G3D and PROC GCONTOUR are then employed to produce three dimensional plots and contour plots of the appropriate function. Output from these procedures appear in the next chapter. The FORTRAN routine CPLOT written by Phil Spector provides a contour printer plot. Output from this routine appears following the parameter estimates for each order

of approximation. The use of CPLOT has currently been suppressed because of the availability of SAS/GRAPH.

5.5 A Note on Computer Facilities

The programs mentioned were developed and run on an Amdahl 470 V/6 or Amdahl 470 V/78 operated by the Data Processing Center at Texas A and M University. The operating system employed is MVS/JES3 allowing the joint operation of both Amdahls as a single system. The WYLBUR text manipulation system was used to type, edit, and run the programs from a CRT terminal.

The program BISAM is executed using a one step system procedure FORTX that compiles, loads, and executes a FORTRAN program. Since BISAM has been in the developmental stage, a compiled version has not been created so run times reflect the more inefficient procedure employed. A typical run of BISAM for a bivariate data set of size 100 requesting all output will use about 50 CPU seconds and will entail the reading of 2398 card images. Naturally, different data sets of the same size may have vastly different characteristics (such as the number of tied observations, correlation, etc.) so that the above represents only an approximation. Furthermore, the above results are by no means typical for all computers and are stated here only as a rough guide.

6. EXAMPLES AND APPLICATIONS

6.1 Introduction

Testing a new statistical technique entails two stages of verification:

1) Check the techniques using data from known parent distributions.

2) Expose the techniques to "real data" and compare and contrast the results to those obtained using other techniques.

Poor results at stage (1) should cause one to discard a trial methodology without bothering to continue to stage (2), although valid exceptions to this practice may occur. Often poor results from the first stage suggest modifications to improve a procedure so that it need not be completely discarded. The second stage should be emphasized, however, as it simulates an environment in which a technique will actually be used. One need not confirm the analysis of others. In fact, if the new technique is more sensitive than existing ones, it may suggest additional interpretations of experimental results that are more appealing than those previously obtained. In the area of data analysis, one is especially interested in obtaining as much insight as possible about the nature of a data set. Hence, a technique offering extended diagnostics is especially welcome. One must be careful, however, to avoid being overwhelmed by an abundance

of diagnostics that may be performing similar tasks. Each statistician must choose those techniques which he feels are best suited for data analysis, and if additional procedures are warranted, they may be applied as needed. For example, if a statistician prefers to use two density estimation techniques to get an idea of the distribution of a data set, and both techniques give conflicting results, a third technique may be employed in an attempt to verify or contradict the results already obtained. Consequently, one may prefer to withhold sophisticated and expensive procedures unless the easier, less expensive methods fail to adequately deal with the problem at hand. Many statisticians are content to use a histogram to diagnose symmetry when the inference tool to be employed is fairly robust to slight deviations from symmetry, but rarely would a histogram be adequate to confirm rigid distributional requirements.

In this chapter, some of the procedures developed in previous chapters will be exposed to simulated and real data sets. At stage one, techniques with subjective smoothing requirements are often easily made to conform to the simulated shape or value. It is at stage two that the weaknesses of the subjective factors involved are exposed. Interpretations will be offered for conflicting results.

6.2 Univariate Examples

To examine the univariate density technique of Chapter 3, we consider three data sets. These are given by

A) 100 Gamma(10,1) r.v.'s

B) 50 N(0,1) r.v.'s mixed with 50 N(3,.25) r.v.'s

C) 63 observations on snowfall in Buffalo, New York, from 1910 to 1972.

Listings of these data sets may be found in Tables 5 through 7.

Table 5. 100 Observations from a Gamma(10,1) Distribution

4.24783	7.13944	8.54367	10.02394	12.09786
5.04321	7.15666	8.61258	10.08187	12.12808
5.24307	7.21348	8.62224	10.31604	12.30663
5.42475	7.27734	8.63897	10.42786	12.41536
5.44763	7.32370	8.77373	10.54578	12.53470
5.53802	7.39561	8.87619	10.56661	12.81958
5.59490	7.47524	8.90276	10.57409	12.88938
5.67504	7.50638	9.06354	10.67747	12.96910
6.13585	7.54331	9.09246	10.90295	13.02124
6.20833	7.83998	9.23594	10.90404	13.75874
6.21723	7.84698	9.25439	11.00485	13.87460
6.37002	7.94389	9.37018	11.13315	13.95213
6.38983	8.03307	9.38267	11.13889	14.08677
6.46811	8.03327	9.49000	11.28637	14.19301
6.50198	8.10931	9.51031	11.31890	14.19326
6.52324	8.12584	9.57556	11.38006	15.87417
6.57701	8.19711	9.63029	11.47916	16.10650
6.61291	8.31321	9.81822	11.61629	18.13609
6.75889	8.39989	9.89115	11.69777	18.79675
7.01929	8.40721	9.96447	12.02569	19.88026

Tables 8 through 10 give selected output from the SAS procedure UNIVARIATE for the data sets. Also, to illustrate the utility of the parametric application of minimum information density estimation, we

Table 6. Sample from the Normal Mixture .5 N(0,1) + .5 N(3,.25)

-2.46836	-0.07938	0.93945	2.58292	3.07817
-1.76497	-0.07224	1.04555	2.61084	3.08599
-1.54261	-0.03350	1.15230	2.65899	3.08894
-1.39970	-0.01924	1.15795	2.66611	3.18530
-1.18157	0.02757	1.27020	2.66671	3.19295
-0.92799	0.03663	1.32825	2.69727	3.21383
-0.78408	0.05647	1.35215	2.71167	3.22806
-0.78233	0.08547	1.51836	2.73020	3.23082
-0.72175	0.09712	1.86996	2.76127	3.27329
-0.71528	0.40045	1.91409	2.83441	3.28711
-0.69271	0.54350	1.99246	2.83984	3.34942
-0.63495	0.57335	2.01725	2.84271	3.43836
-0.62522	0.57826	2.17318	2.84482	3.49594
-0.58790	0.64629	2.17619	2.86786	3.49905
-0.42311	0.68756	2.25658	2.92654	3.52823
-0.41908	0.70090	2.28519	2.95894	3.68402
-0.39825	0.79734	2.39207	3.02427	3.70910
-0.22372	0.82677	2.42322	3.03565	3.74198
-0.19536	0.83868	2.43118	3.05925	4.02736
-0.18169	0.90461	2.47505	3.07473	4.17188

will consider a set of 100 simulated N(0,1) random variables and compare the least squares estimates to the usual maximum likelihood or UMVU estimates of the parameters. The normal data set is not exposed to the other procedures because they all seem to perform well for smooth, symmetric densities. Data sets A and B will seek to test the techniques for sensitivity to skewness or bimodality, while data set C is included because it has been analyzed in a variety of density estimation references (see, e.g., Parzen, 1979, or Tapia and Thompson, 1978).

The popularity of the histogram and the fact that the bin width

Table 7. Yearly Snowfall in Buffalo, New York, 1910-1972

25.0	58.0	77.8	85.5	104.5
39.8	60.3	78.1	87.4	105.2
39.9	63.6	78.4	88.7	110.0
40.1	65.4	79.0	89.6	110.5
46.7	66.1	79.3	89.8	110.5
49.1	69.3	79.6	89.9	113.7
49.6	70.9	80.7	90.9	114.5
51.1	71.4	82.4	97.0	115.6
51.6	71.5	82.4	98.3	120.5
53.5	71.8	83.0	101.4	120.7
54.7	72.9	83.6	102.4	124.7
55.5	74.4	83.6	103.9	126.4
58.0	77.8	85.5		

problem is well documented make its analysis here unnecessary. For illustrative purposes, we include Figure 6 showing the output of PROC CHART of SAS for the Buffalo snowfall data.

Using the FORTRAN routines documented in the last Chapter, we obtain density estimates to be labeled as follows:

FNN - Nearest Neighbor density estimate

FKR - Kernel estimate

FKT - Kronmal-Tartar trigonometric series estimate

FTK - Tartar-Kronmal complex Fourier series estimate

FMI - minimum information estimate using Legendre [-1,1] polynomials

FMC - minimum information estimate using complex Fourier series

Table 8. Selected Output from SAS PROC UNIVARIATE for Data Set A

MOMENTS

N	100	SUM WGTS	100
MEAN	9.64664	SUM	964.664
STD DEV	3.06639	VARIANCE	9.40273
SKEWNESS	0.904072	KURTOSIS	1.05021
USS	10236.6	CSS	930.871
CV	31.7871	STD MEAN	0.306639
T:MEAN=0	31.4593	PROB>|T|	0.0001

QUANTILES

100% MAX	19.8802	99%	19.8694
75% Q3	11.3648	95%	15.7901
25% Q1	7.34168	10%	6.20922
0% MIN	4.24783	5%	5.45215
		1%	4.25578
RANGE	15.6324		
Q3-Q1	4.0231		
MODE	4.24783		

FTK will only be applied to data set C to illustrate the MISE order determining criterion. Examples of the application of autoregressive density estimation to Buffalo snowfall may be found in Parzen (1979b) along with other examples, and hence will not be included here. Tapia and Thompson (1978) also give a variety of examples for various density estimation routines. Tartar and Kronmal (1976) illustrate the use of FTK applied to relatively smooth data sets. We include the above estimates for comparison and illustrative purposes.

For data set A, Figures 7 through 11 contain the "best" estimates obtained from a procedure overlayed with the true population density

Table 9. Selected Output from SAS PROC UNIVARIATE for Data Set B

MOMENTS

N	100	SUM WGTS	100
MEAN	1.52001	SUM	152.001
STD DEV	1.62488	VARIANCE	2.64023
SKEWNESS	-0.359466	KURTOSIS	-1.07955
USS	492.427	CSS	261.383
CV	106.899	STD MEAN	0.162488
T:MEAN=0	9.35463	PROB>|T|	0.0001

QUANTILES

100% MAX	4.17188	99%	4.17043
75% Q3	2.95084	95%	3.67623
50% MED	1.95327	90%	3.34319
25% Q1	0.029835	10%	-0.713023
0% MIN	-2.46836	5%	-1.16889
		1%	-2.46133
RANGE	6.64024		
Q3-Q1	2.921		
MODE	-2.46836		

used to generate the data. One notes that FNN has difficulty smoothing out the mode of the density, but the value at the mode is 0.14 which is close to the true value 0.13. The densities FKR, FMI, and FMC all perform an adequate approximation to the parent gamma(10,1) density. Coefficients for FKT for the three data sets are displayed in Table 11. Table 12 contains the coefficients for FMI, and Table 13 contains coefficients for FMC.

For data set B, some of the techniques have a little more difficulty approximating the bimodal parent density. Figures 12

Table 10. Selected Output from SAS PROC UNIVARIATE for Data Set C

MOMENTS

N	63	SUM WGTS	63
MEAN	80.2952	SUM	5058.6
STD DEV	23.7198	VARIANCE	562.629
SKEWNESS	-0.0186313	KURTOSIS	-0.562101
USS	441065	CSS	34883
CV	29.5407	STD MEAN	2.98842
T:MEAN=0	26.8688	PROB>|T|	0.0001

QUANTILES

100% MAX	126.4	99%	126.4
75% Q3	98.3	95%	120.66
50% MED	79.6	90%	114.18
25% Q1	63.6	10%	49.3
0% MIN	25	5%	39.94
		1%	25
RANGE	101.4		
Q3-Q1	34.7		
MODE	82.4		

through 17 contain plots for this data set. FNN once again seems to fluctuate randomly about the modes but gives a good rough approximation to the bimodal shape. FKR, FMI, and FMC once again seem to provide the best results with some problem in estimating the true value of the parent density at the modes, but diagnosing the bimodal shape well. Figure 17 shows the exceptional ability when presented with an above average initial density estimate in CMPDEN instead of FNN, and the resulting value of FMC is extremely close to the true values of the parent density. The only discrepancies occur in the

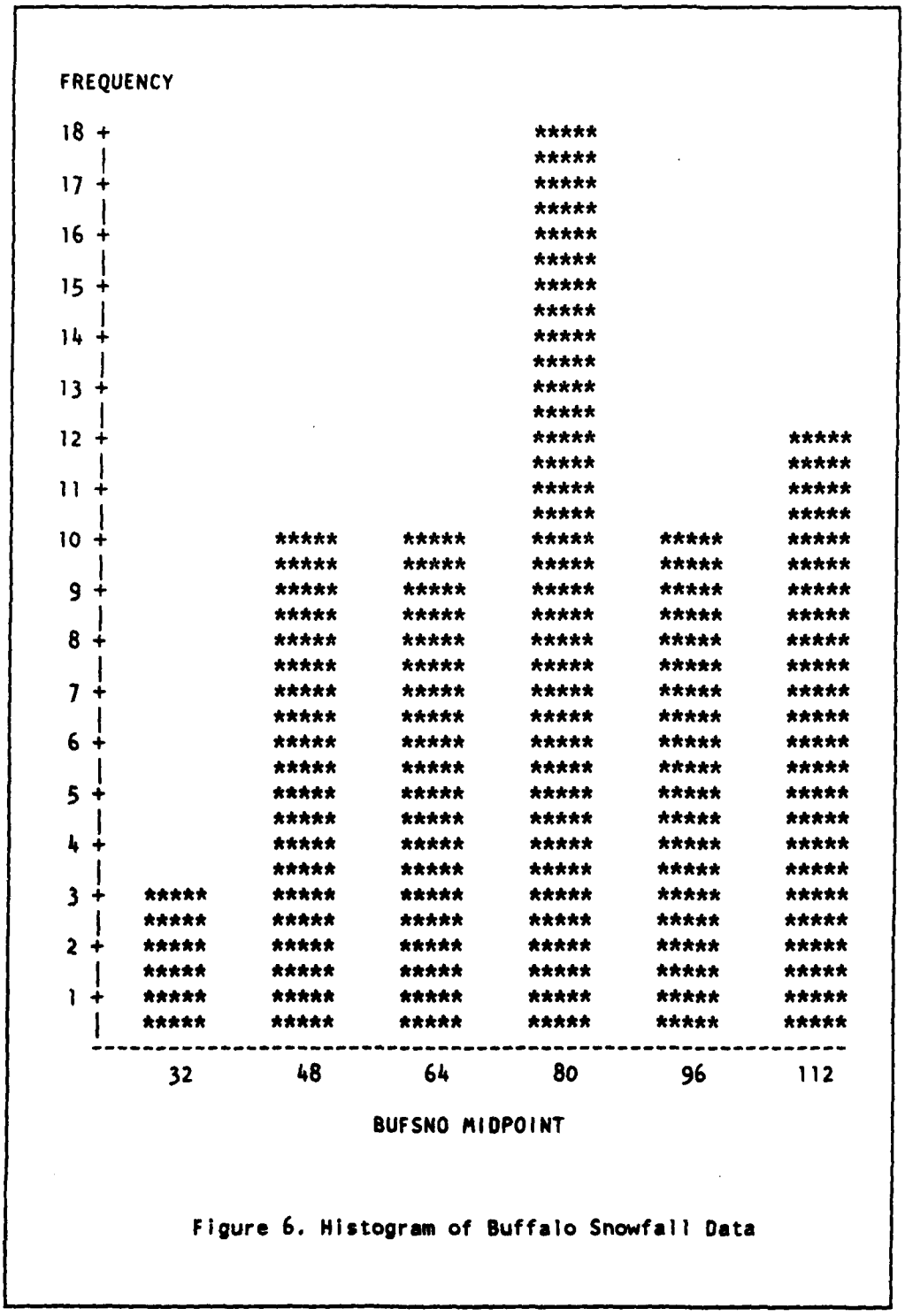


Figure 6. Histogram of Buffalo Snowfall Data

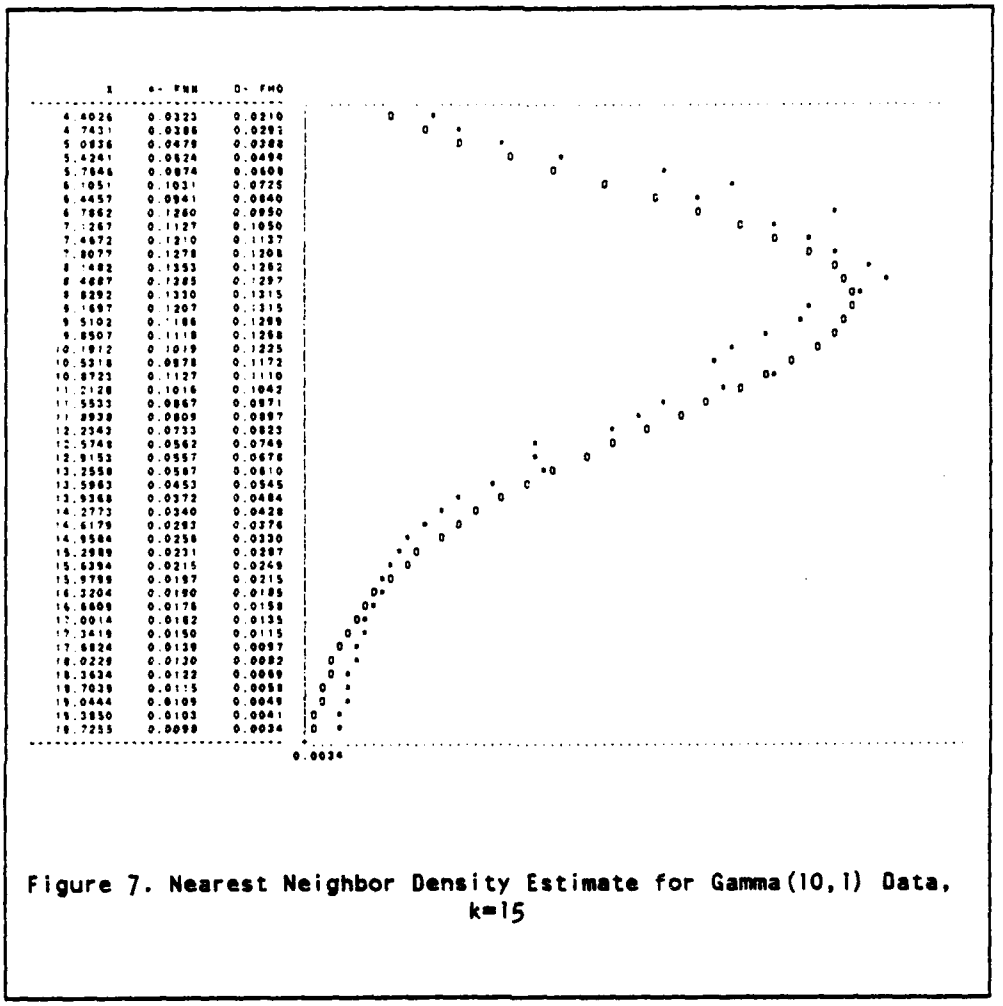


Figure 7. Nearest Neighbor Density Estimate for Gamma(10,1) Data, k=15

tail areas which is typical for the orthogonal expansion techniques.

For data set C, Figures 18 through 23 represent the different shapes one may subjectively obtain from a density estimation procedure. FNN seems to indicate a trimodal parent density. FKR depicts a unimodal shape. FTK is included in this analysis with coefficients listed in Table 14. For FKT and FTK, functionals of the estimated p.d.f. were included to compare to the usual unbiased

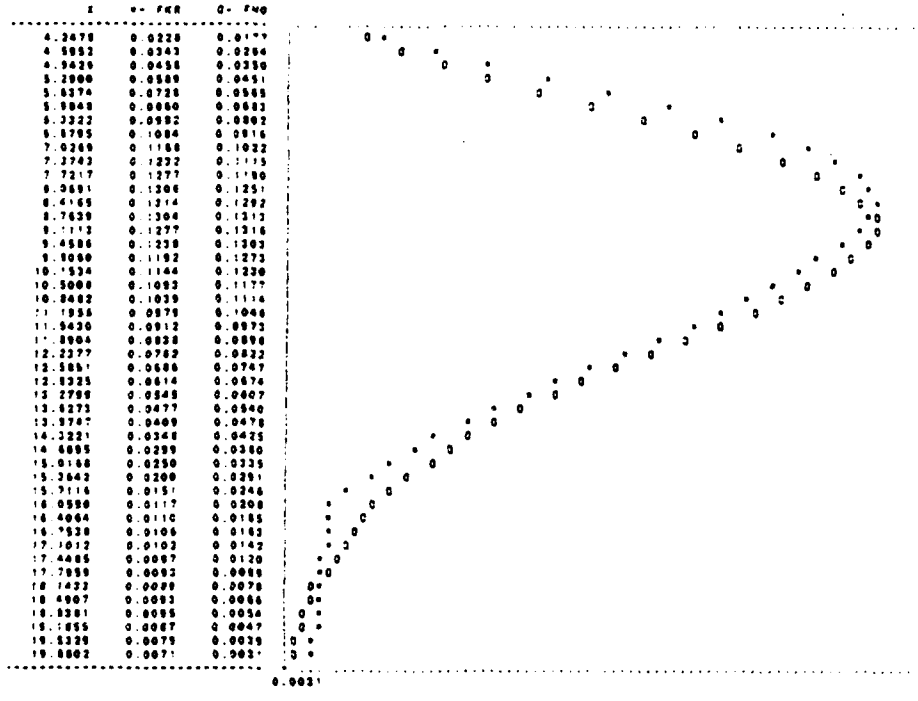


Figure 8. Kernel Density Estimate for Gamma(10,1) Data, h=1.0

estimates of location and scale. Table 15 contains selected functionals of the estimated densities for the three data sets. Recall from Table 10 (p. 160) the sample mean for Buffalo snowfall is 80.29 and the sample variance is 562.629. Typically, for different shapes, an orthogonal expansion estimate presents fairly stable functional values for the mean and variance. FKT and FTK admit both unimodal and trimodal shapes. Using the MISE criterion for FTK, a

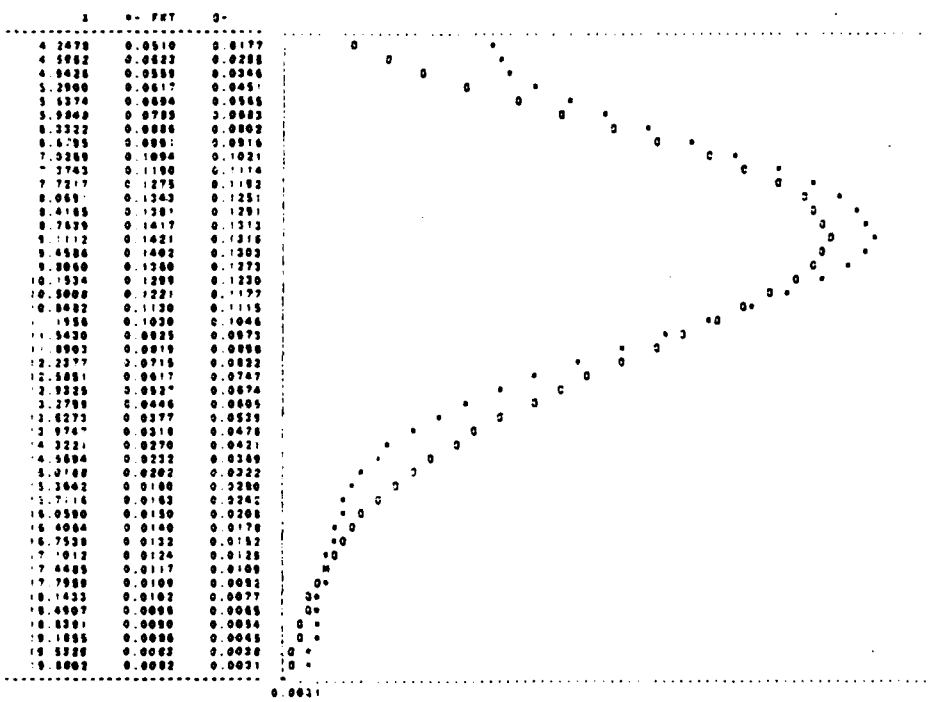


Figure 9. Trigonometric Series Density Estimate for Gamma(10,1) Data, m=7

unimodal shape is obtained with mean and variance "close" to their unbiased counterparts. An order 7 approximation for FKT is displayed to illustrate a trimodal shape, although a unimodal shape is obtained for orders 1 through 4. Again, the mean and variance estimates are similar to their unbiased counterparts.

FMI and FMC tend to provide different shapes, and as exemplified above, the complex expansion tends to introduce multimodal estimates

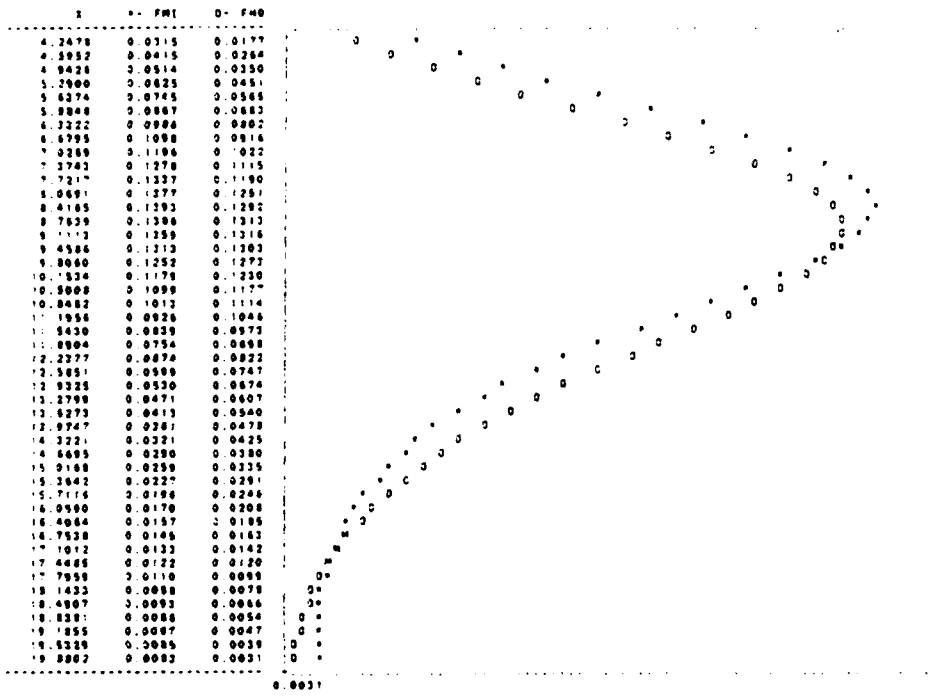


Figure 10. Minimum Information Series Estimate using Legendre Polynomials for Gamma(10,1) Data, m=3

at lower orders than the polynomial expansions. The order 6 estimate for FMI depicts the transition from a unimodal to a trimodal shape. The order 6 estimate for FMC already clearly indicates a trimodal shape.

One might note that for the two objective criteria MISE and CAT, a unimodal shape for Buffalo snowfall is indicated, although most procedures will admit trimodal shapes. (See Parzen, 1979, for

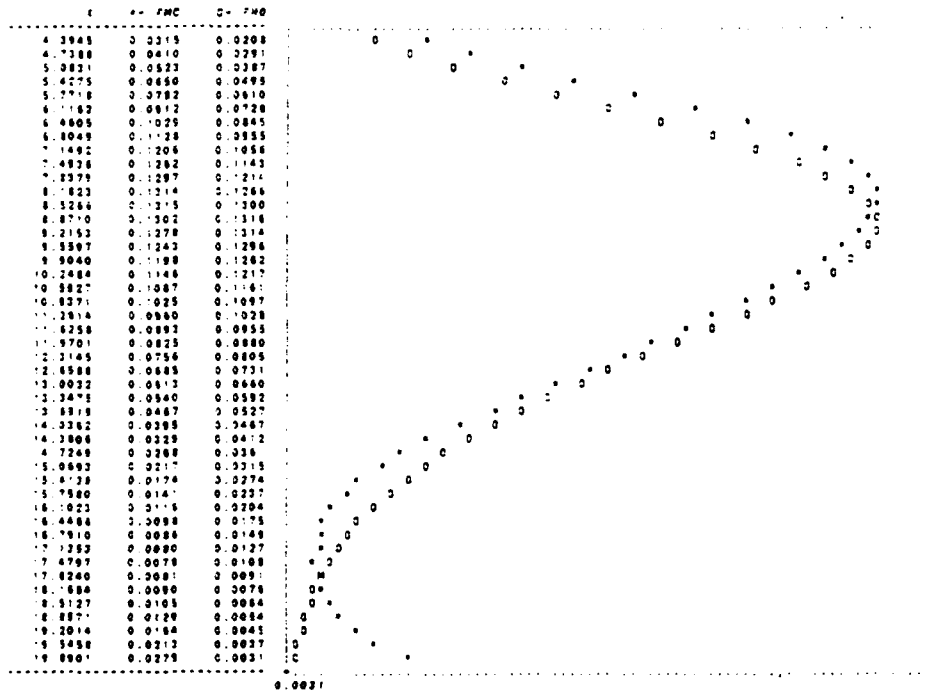


Figure 11. Minimum Information Series Estimate using Complex Exponentials for Gamma(10,1) Data, m=6

illustration of CAT results). When the true p.d.f. is known, as is the case for data sets A and B, a procedure can usually be steered to produce a desired shape. Consequently, a better test for a procedure would be for stability or objectivity. The estimate FNN is fairly stable for orders of 5 or more in terms of representing a fixed number of modes although fluctuations in mode estimates occur. The orthogonal expansion techniques tend to be unstable, especially for

Table 11. Trigonometric Series Coefficients

DATA SET A: I	COF	DATA SET B: I	COF
1	0.0530	1	-0.0752
2	-0.0234	2	-0.0235
3	-0.0317	3	-0.0289
4	-0.0111	4	-0.0595
5	-0.0123	5	0.0779
6	-0.0068	6	-0.0456
7	-0.0044	7	0.0393
8	-0.0027	8	0.0123
9	-0.0021	9	-0.0222
10	-0.0015	10	-0.0071

DATA SET C: I	COF
1	-0.0021
2	-0.0057
3	-0.0002
4	0.0014
5	-0.0023
6	-0.0021
7	0.0032
8	0.0014
9	-0.0013
10	0.0006

large orders of approximation, but objective use of criterion functions helps to overcome this problem. The MISE criterion of Tartar and Kronmal (1970,1976) does not translate effectively to the *minimum information procedure*, but one suspects a modification of MISE, CAT, or AIC should better handle the problem. Further research is warranted in this area. Consequently, we do not advocate as yet the use of minimum information techniques over autoregressive estimation with the CAT criterion or estimation using FTK with the

Table 12. Coefficients for FMI Density Estimate

Data Set A:	Order	Coef	Data Set B:	Order	Coef
	1	-1.411		1	0.658
	2	-0.990		2	-0.986
	3	0.749		3	0.228
				4	-0.746
Data Set C:	1	0.237		5	-0.977
	2	-0.951		6	0.267
	3	0.012		7	0.552
	4	0.164		8	0.392
	5	0.082			
	6	-0.227			

MISE criterion. However, in the bivariate case, the minimum information approach is easily extended and seems more appropriate than existing procedures, especially in terms of controlling the amount of graphical displays necessary to arrive at an acceptable estimate.

To illustrate the parametric application of the minimum information approach to estimating parameters for normal and gamma models, data set A is examined along with data set D consisting of a random sample of 100 N(0,1) values. A parametric model for data set A appears in Figure 24. Note the least squares estimates are a=8.94 and b=0.94 which correspond to an estimated gamma mean of 9.51 and an estimated gamma variance of 10.12. Data set D is listed in Table 16 with selected descriptive statistics from PROC UNIVARIATE of SAS

Table 13. Coefficients for FMC Density Estimate

DATA SET A:	IND	REAL (THETA)	IMAG (THETA)
	-1	-0.316603	0.596658
	1	-0.316603	-0.596659
	-2	-0.070131	0.161096
	2	-0.070130	-0.161097
	-3	-0.056627	0.069507
	3	-0.056627	-0.069507

DATA SET B:	IND	REAL (THETA)	IMAG (THETA)
	-1	-0.175565	-0.197545
	1	-0.175564	0.197546
	-2	-0.204320	-0.293631
	2	-0.204320	0.293632
	-3	-0.114078	-0.043741
	3	-0.114078	0.043741

DATA SET C:	IND	REAL (THETA)	IMAG (THETA)
	-1	-0.351673	-0.079655
	1	-0.351673	0.079656
	-2	-0.065335	-0.074402
	2	-0.065334	0.074402
	-3	-0.117258	-0.186600
	3	-0.117258	0.186600

appearing in Table 17. Figure 25 shows the parametric representation of the normal density with a least squares estimated mean of -0.10 and variance of 1.25. These examples illustrate the parametric applications of the minimum information procedures, suggesting possible extensions to goodness-of-fit diagnostics.

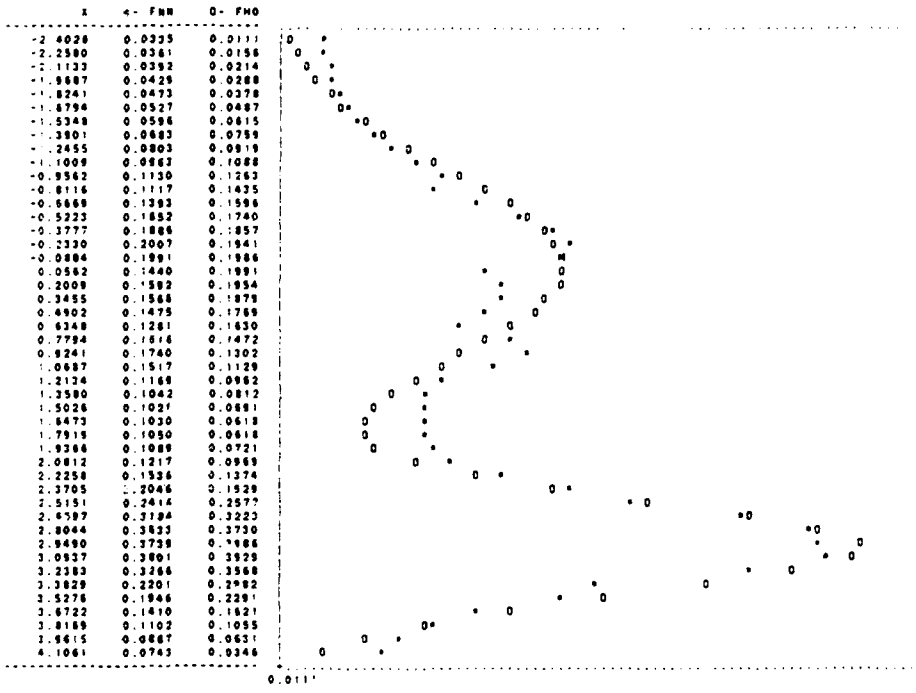


Figure 12. Nearest Neighbor Density Estimate for Normal Mixture Data, k=15

A few comments are in order. Examination of the expansion coefficients reveals a rather rapid decay, as expected, but higher order coefficients usually remain large enough to induce a very wiggly shape to the estimated density. Furthermore, expansions using complex exponentials seem to require fewer terms than the polynomial or trigonometric series expansions. This indicates that complex Fourier series may converge more rapidly than other series expansions, a

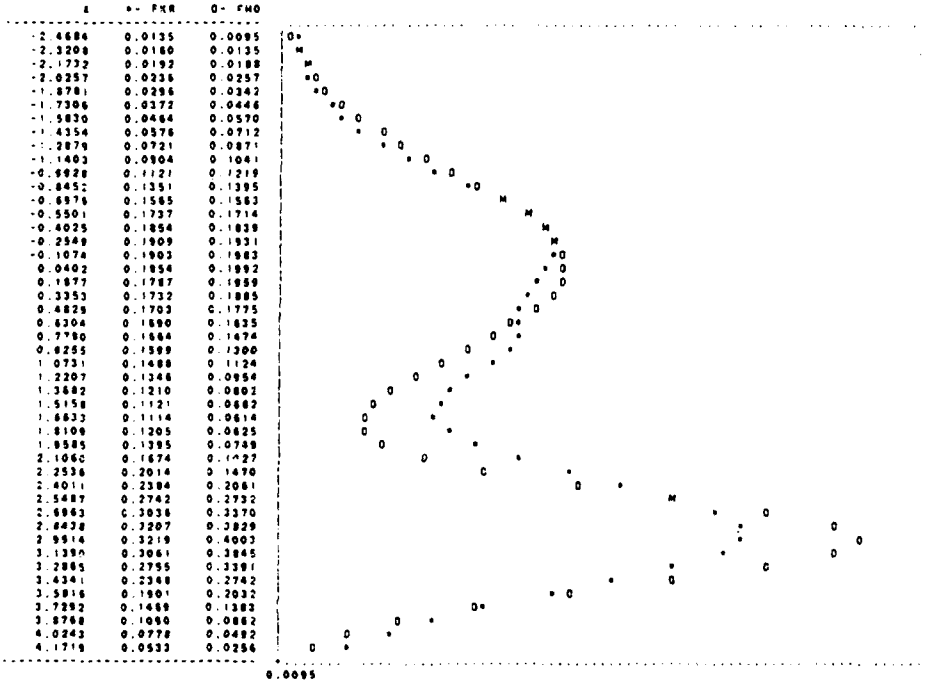


Figure 13. Kernel Density Estimate for Normal Mixture Data, h=1.0

conjecture that seems to be supported by the literature. This motivated using only complex expansions in the bivariate case, although one may wish to consider other choices for the orthogonal system of functions to use.

In this section we have presented several examples of nonparametric density estimation procedures. Essentially, we have let the plots speak for themselves to illustrate obtainable shapes and the

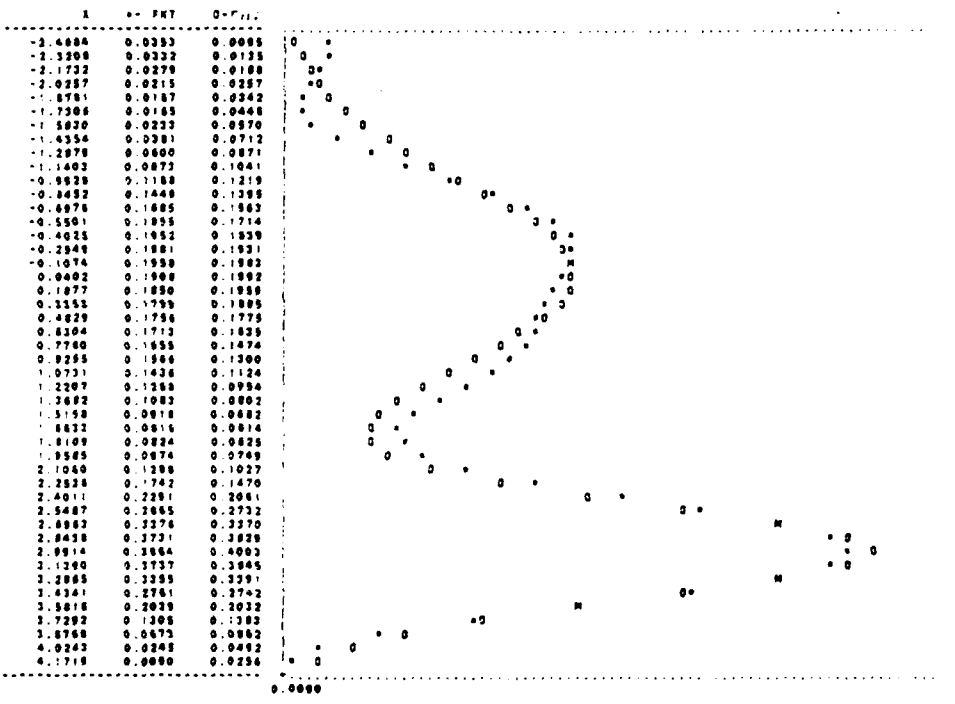


Figure 14. Trigonometric Series Density Estimate for Normal Mixture Data, m=6

corresponding anomalies that are produced. The examination of the univariate MIDEN and CMPDEN procedures illustrates the power of the minimum information approach and justifies the examination of bivariate extensions. Had CMPDEN performed poorly in the univariate case, one would have little faith in the ability of its bivariate counterpart. With this motivation, one may now examine the bivariate extension of CMPDEN.

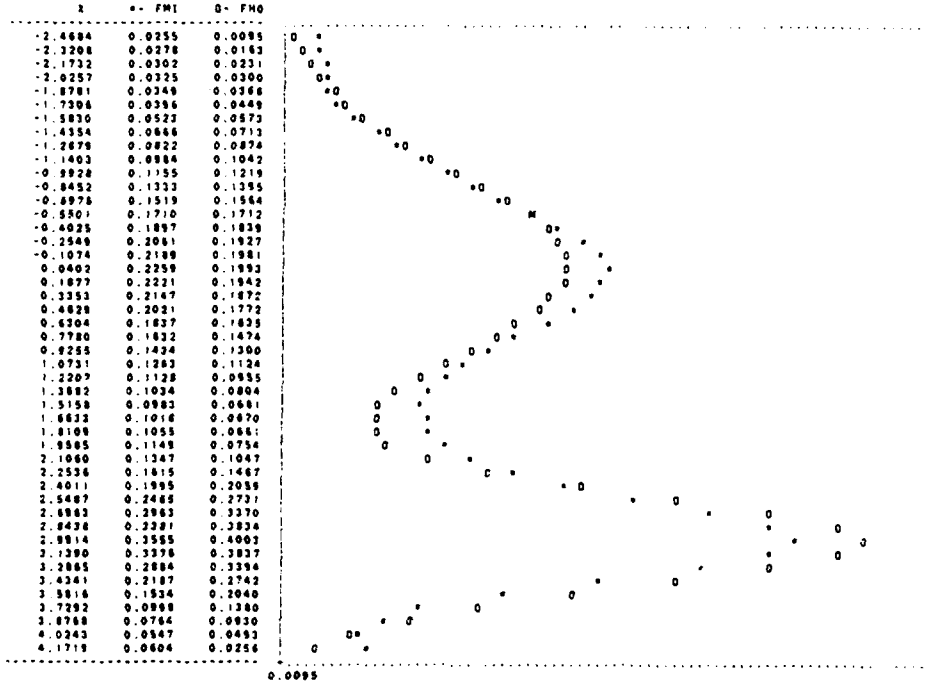


Figure 15. Minimum Information Series Estimate using Legendre Polynomials for Normal Mixture Data, m=8

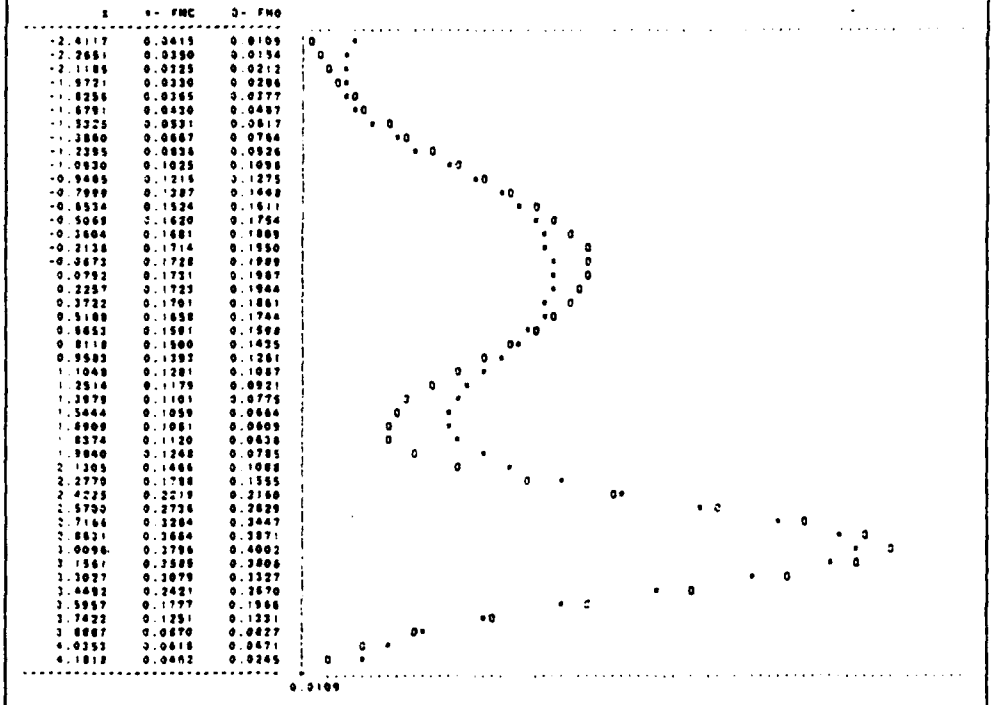


Figure 16. Minimum Information Series Estimate using Complex Exponentials for Normal Mixture Data, m=6

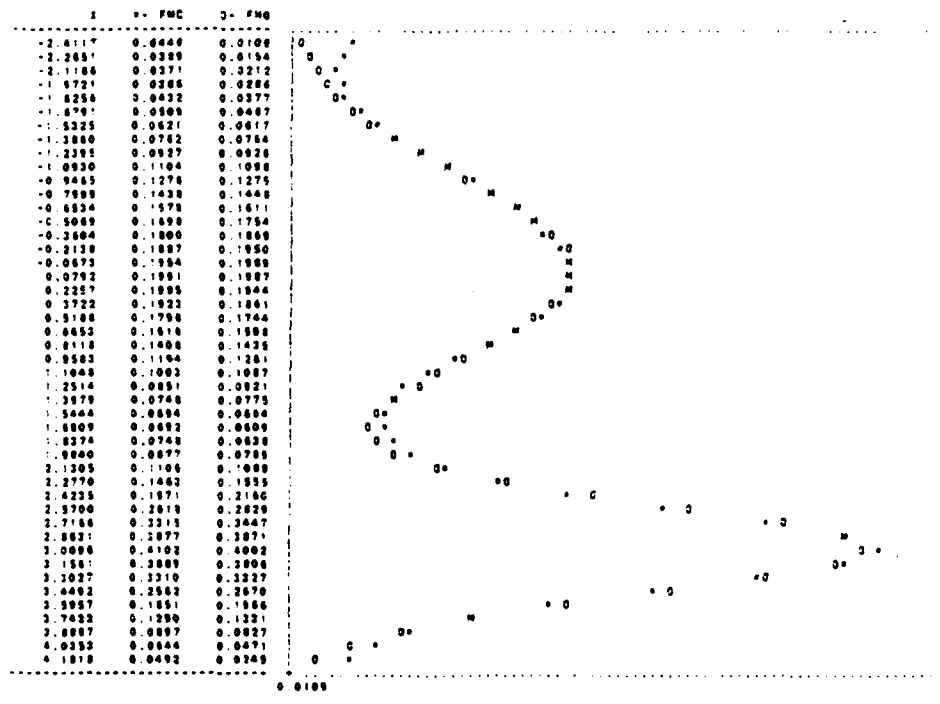


Figure 17. Minimum Information Series Estimate using Complex Exponentials with Kernel Initial Estimate for Normal Mixture Data, m=6

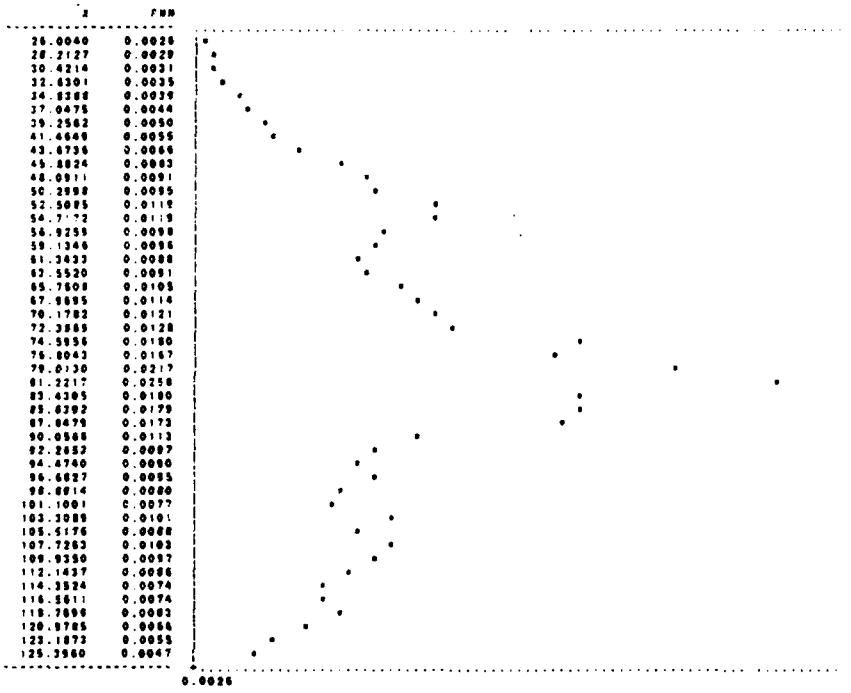


Figure 18. Nearest Neighbor Density Estimate for Buffalo Snowfall Data, k=10

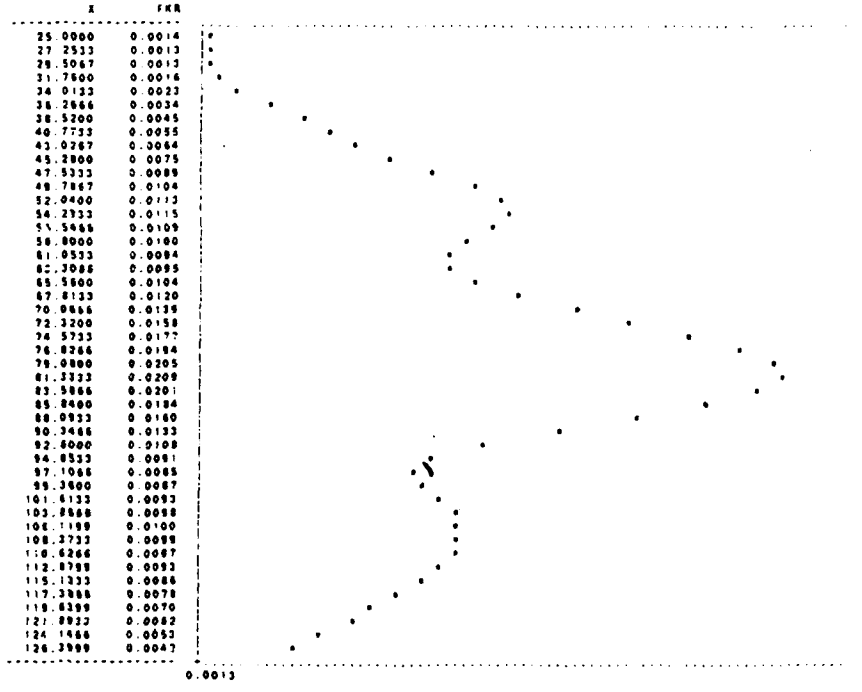


Figure 19. Kernel Density Estimate for Buffalo Snowfall Data, h=5.0

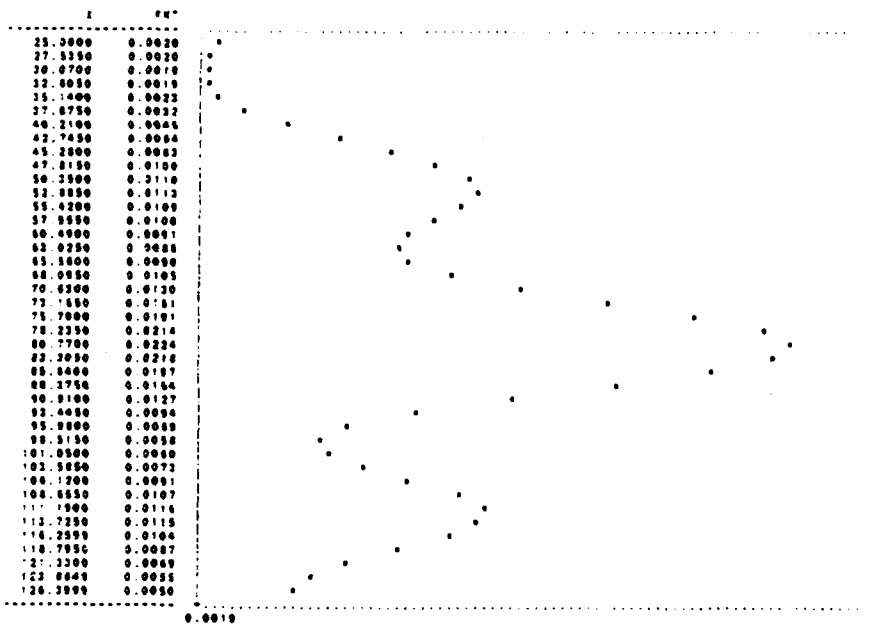


Figure 20. Trigonometric Series Density Estimate for Buffalo Snowfall Data, m=7

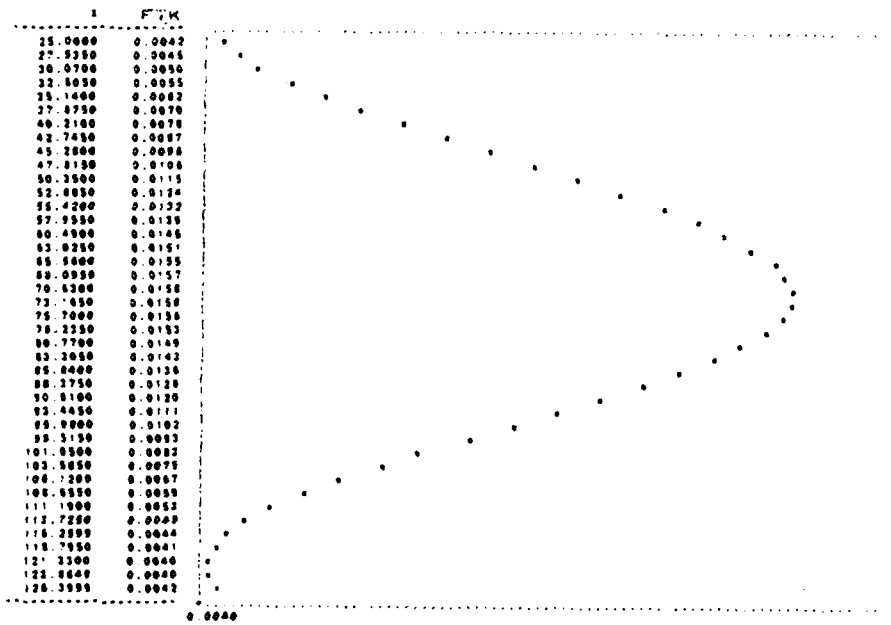
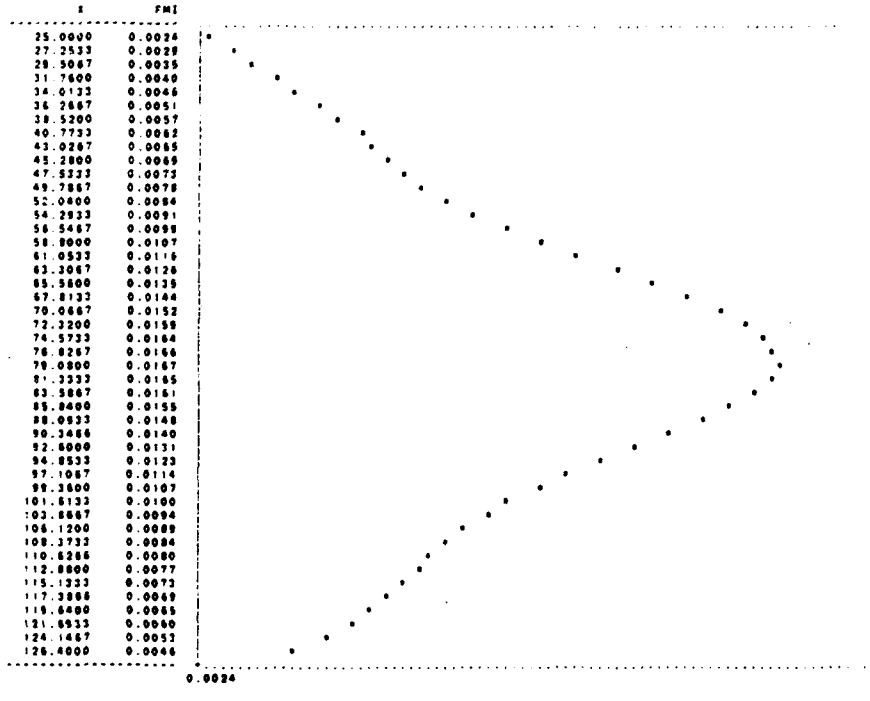


Figure 21. Complex Exponential Series Density Estimate for Buffalo Snowfall Data, m=1



I	FMI
25.0000	0.0024
27.2533	0.0028
29.5067	0.0035
31.7600	0.0040
34.0133	0.0046
36.2667	0.0051
38.5200	0.0057
40.7733	0.0062
43.0267	0.0065
45.2800	0.0069
47.5333	0.0073
49.7867	0.0078
52.0400	0.0084
54.2933	0.0091
56.5467	0.0098
58.8000	0.0107
61.0533	0.0116
63.3067	0.0126
65.5600	0.0135
67.8133	0.0144
70.0667	0.0152
72.3200	0.0159
74.5733	0.0164
76.8267	0.0166
79.0800	0.0167
81.3333	0.0165
83.5867	0.0161
85.8400	0.0155
88.0933	0.0148
90.3466	0.0140
92.6000	0.0131
94.8533	0.0123
97.1067	0.0114
99.3600	0.0107
101.6133	0.0100
103.8667	0.0094
106.1200	0.0089
108.3733	0.0084
110.6266	0.0080
112.8800	0.0077
115.1333	0.0073
117.3866	0.0069
119.6400	0.0065
121.8933	0.0060
124.1467	0.0053
126.4000	0.0046

Figure 22. Minimum Information Series Estimate using Legendre Polynomials for Buffalo Snowfall Data, m=6

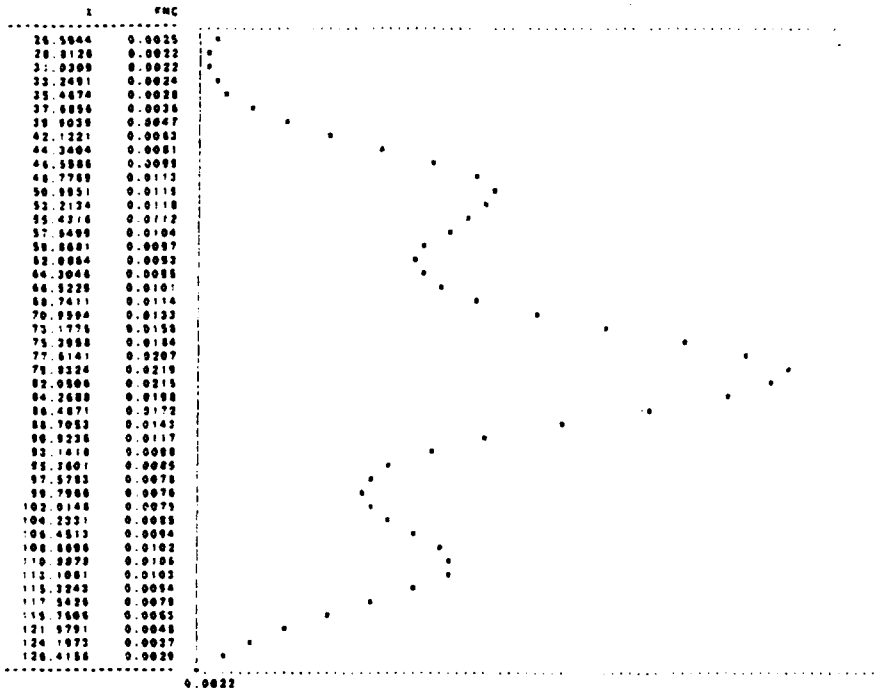


Figure 23. Minimum Information Series Estimate using Complex Exponentials for Buffalo Snowfall Data, m=6

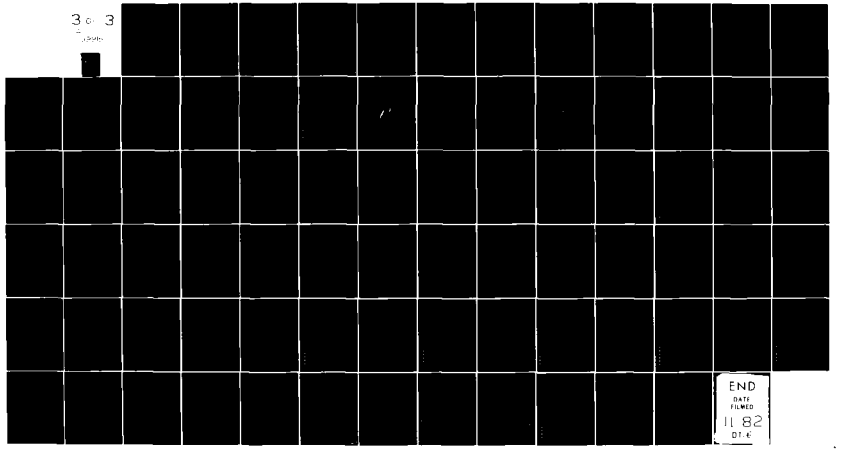
Table 14. Coefficients for FTK Density Estimate

DATA SET C:	IND	REAL(THETA)	IMAG(THETA)
	-1	-0.2881	-0.0793
	1	-0.2881	0.0793
	-2	0.0698	-0.0014
	2	0.0698	0.0014
	-3	-0.1041	-0.1915
	3	-0.1041	0.1915
	-4	0.0732	0.0980
	4	0.0732	-0.0980
	-5	0.0318	-0.0290
	5	0.0318	0.0290
	-6	-0.0218	-0.0974
	6	-0.0218	0.0974

Table 15. Selected Functionals of Estimated Densities

Estimate	Mean	Variance	ISE	MSE	Max. Dev.
Data Set A:					
True	10.00	10.00	-	-	-
FNN (k=15)	9.79	10.25	0.0024	0.0002	0.0352
FKT (m=7)	9.65	9.47	0.0019	0.0001	0.0332
FMI (m=3)	9.48	9.01	0.0025	0.0002	0.0185
Param. L.S.	9.51	10.12	0.0015	0.0001	0.0205
Data Set B:					
True	1.50	2.88	-	-	-
FNN (k=15)	1.49	2.79	0.0055	0.0009	0.0781
FKR (h=0.4)	1.50	2.61	0.0077	0.0016	0.0805
FKT (m=6)	1.53	3.28	0.0023	0.0002	0.0343
FMI (m=8)	1.31	2.79	0.0068	0.0018	0.0553
Data Set C:					
FNN (k=10)	78.90	548.77	-	-	-
FKR (h=5.)	79.92	524.36	-	-	-
FKT (m=7)	80.33	548.67	-	-	-
FTK (m=1)	73.15	558.96	-	-	-
FMI (m=6)	76.79	595.74	-	-	-
FMC (m=6)	79.03	532.64	-	-	-
Data Set D:					
True	0.0	1.0	-	-	-
Param. L.S.	-0.0032	1.04	0.0001	0.0000	0.0066

AD-A119 915 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
STATISTICAL MODELING OF BIVARIATE DATA.(U)
AUG 82 T J WOODFIELD
UNCLASSIFIED TR-8-7 DAAG29-80-C-0070
ARO-16992.13-MA NL



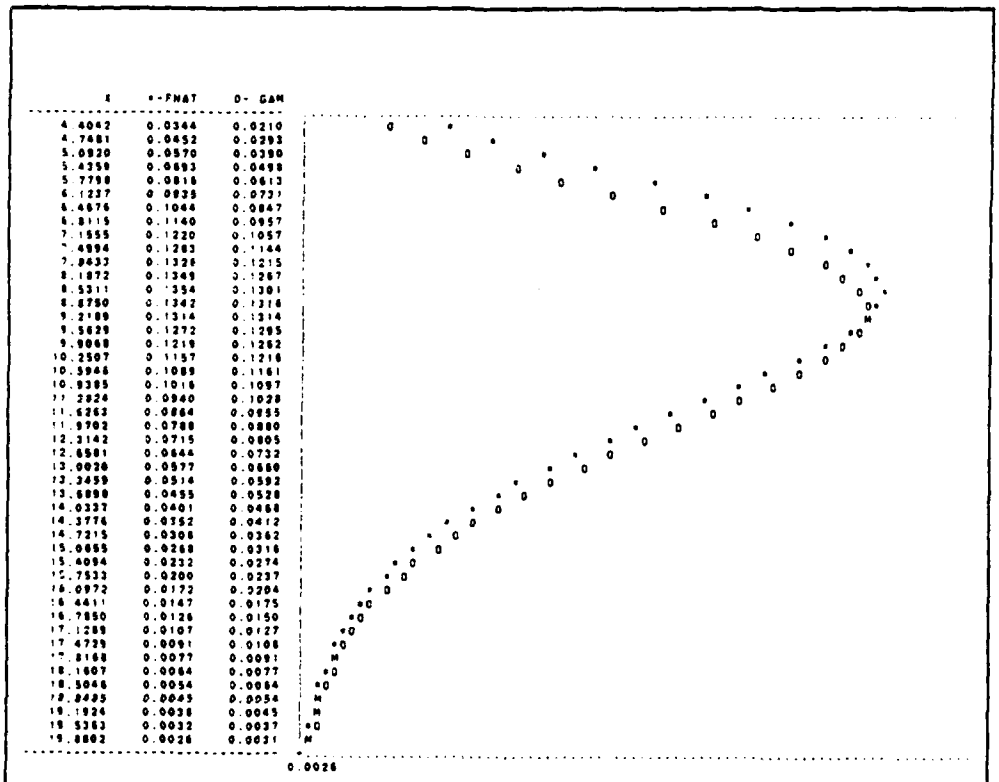


Figure 24. Parametric Gamma Least Squares Density Estimate

Table 16. 100 Observations from a Normal(0,1) Distribution

-2.4543	-0.9761	-0.2562	0.2586	0.8609
-1.9855	-0.9257	-0.2454	0.3443	0.9585
-1.9848	-0.8681	-0.2044	0.3506	1.0888
-1.7842	-0.8600	-0.1885	0.3836	1.0903
-1.5225	-0.8447	-0.1862	0.3875	1.1151
-1.4707	-0.8020	-0.1796	0.4137	1.1342
-1.4021	-0.7875	-0.1696	0.4164	1.2297
-1.3537	-0.7511	-0.1534	0.4279	1.2326
-1.3480	-0.7226	-0.1420	0.4341	1.3730
-1.2972	-0.6578	-0.1235	0.4372	1.4012
-1.2791	-0.6520	-0.0955	0.4471	1.4153
-1.2261	-0.5746	-0.0321	0.4938	1.4422
-1.2138	-0.5337	0.0161	0.5047	1.4573
-1.2040	-0.5241	0.0852	0.5215	1.4895
-1.1369	-0.4537	0.1382	0.6104	1.5830
-1.0814	-0.4434	0.1409	0.6427	1.6441
-1.0793	-0.4220	0.1521	0.6830	1.6826
-1.0661	-0.3949	0.1623	0.7354	1.7450
-1.0071	-0.3725	0.1697	0.7788	1.8024
-0.9990	-0.2596	0.2132	0.7833	2.3243

Table 17. Selected Output from SAS PROC UNIVARIATE for Data Set D

MOMENTS

N	100	SUM WGTS	100
MEAN	-0.03496	SUM	-3.496
STD DEV	0.996355	VARIANCE	0.992724
SKEWNESS	0.0461974	KURTOSIS	-0.555495
USS	98.4019	CSS	98.2797
CV	-2849.99	STD MEAN	0.0996355
T:MEAN=0	-0.350879	PROB>|T|	0.726425

QUANTILES

100% MAX	2.3243	99%	2.31908
75% Q3	0.634625	95%	1.64104
50% MED	-0.1095	90%	1.41389
25% Q1	-0.834025	10%	-1.29539
0% MIN	-2.4543	5%	-1.51991
		1%	-2.44961
RANGE	4.7786		
Q3-Q1	1.46865		
MODE	-2.4543		

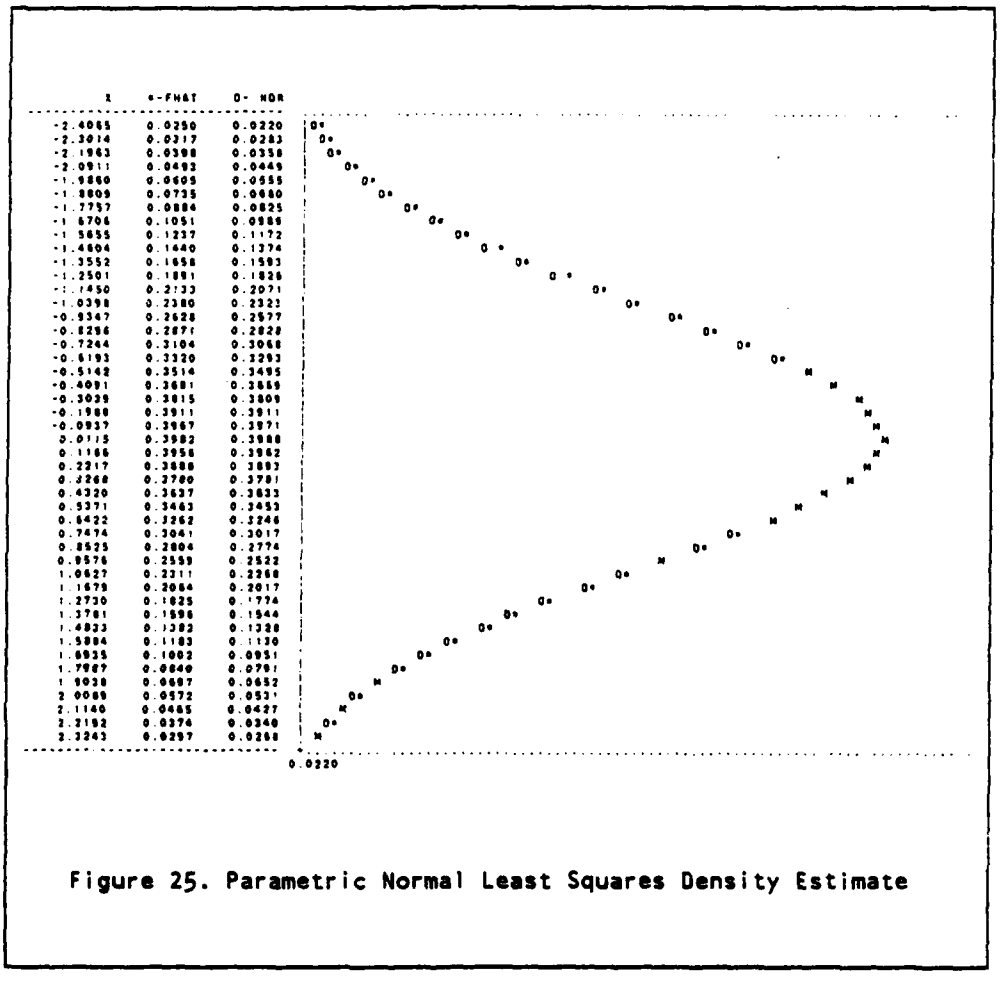


Figure 25. Parametric Normal Least Squares Density Estimate

6.3 Bivariate Examples

To illustrate the use of the BISAM program that implements the bivariate data modeling approach detailed in Chapter 4, we consider three data sets. These are

E) 500 independent N(0,1) bivariate observations,

F) 100 bivariate standard normal variables with a correlation of 0.90

G) The coronary data of Scott, et al. (1978).

Selected output from BISAM for data sets E and F appears in Table 18. Listings of the Coronary data appear in Tables 19 and 20, and selected output from BISAM for this data appears in Tables 21 and 22.

For the normal data we shall merely exhibit the results and note the similarities of the population quantities being estimated. Figures 26 through 29 display contour and three dimensional plots from SAS/GRAPH. Scatter plots have been omitted as have the univariate autoregressive density plots which were normal in all cases. The figures illustrate slight anomalies that may occur in the "off diagonal" areas for a high degree of correlation. These are due in part to the extrapolation problem. Clearly few data points occur in these tail areas to adequately estimate the bivariate density there. For the most part, the bivariate estimates of the normal densities are pleasing. One notes that the mode occuring at the point (-0.31,-0.14) for data set E is a little unusual, but this seems to result from the

Table 18. Summary of BISAM Output for Data Sets E and F

	DATA SET E		DATA SET F	
	X	Y	X	Y
Mean	-0.092	-0.085	-0.026	-0.025
Median	-0.131	-0.145	0.027	0.060
Trimean	-0.109	-0.111	-0.020	0.028
Variance	0.973	0.954	1.120	1.078
St. Dev.	0.987	0.977	1.058	1.038
IQ Range	1.337	1.420	1.378	1.463
Pearson r	-0.027		0.893	
Spearman Rho	-0.038		0.884	
Kendall Tau	-0.026		0.710	
H(d-tilda)	-0.292		-0.858	
H(d8)	-0.056		-0.351	
H(d24)	-0.077		-0.462	
H(d48)	-0.104		-2.621	

simulation and not the modeling technique. The entropy estimates exhibit the same type of instability discussed before indicating that some correction factor should possibly be employed in their computation. The results for data set E with n=500 indicate that the entropy statistics may be asymptotically biased. Otherwise, the results for the normal case are satisfactory, leading one to consider investigations with real data.

Scott, et al. (1978), consider two sets of data consisting of measurements of plasma cholesterol (CHL) concentration and plasma triglyceride (TRG) concentration in 371 males. The males were classified into two groups, 320 falling into the category of "diseased" and 51 being classified as "normal". These classifications

Table 19. Listing of Coronary Data - Normal

PLASMA CHOLESTEROL - NO CORONARY ARTERY DISEASE

195. 237. 205. 201. 190. 180. 193. 170. 150. 200. 228. 169. 178.
251. 234. 222. 116. 157. 194. 130. 206. 158. 167. 217. 234. 190.
178. 265. 219. 265. 190. 156. 187. 149. 147. 155. 207. 238. 168.
210. 208. 160. 243. 209. 221. 178. 289. 201. 168. 162. 207.

PLASMA TRIGLYCERIDES - NO CORONARY ARTERY DISEASE

348. 174. 158. 171. 85. 82. 210. 90. 167. 154. 119. 86. 166.
211. 143. 284. 87. 134. 121. 64. 99. 87. 177. 114. 116. 132.
157. 73. 98. 485. 108. 126. 109. 146. 95. 48. 195. 172. 71.
91. 139. 116. 101. 97. 156. 116. 120. 72. 100. 227. 160.

were based on medical examination of the patients to ascertain the presence of coronary artery disease. Using a Kolmogorov-Smirnov goodness-of-fit test, a null hypothesis of normality is rejected for all but the diseased class of triglyceride data. This motivated the use of the kernel method of bivariate density estimation to assist in the analysis of the data. A likelihood equation was then developed to aid in patient classification and to estimate the risk of coronary artery disease based on CHL and TRG measurements.

It was determined that the normal data exhibited a unimodal shape with mode at (CHL,TRG) = (195,122). The diseased population was felt to exhibit a bimodal shape with one mode at M1 = (185,122) and the second mode at M2 = (233,145). The univariate kernel estimates were unimodal. Consequently, it was felt that some diseased patients were virtually indistinguishable from normal patients based on CHL and TRG

Table 20a. Listing of Coronary Data - Diseased

PLASMA CHOLESTEROL - DISEASE IN AT LEAST 1 OF 3 CORONARY ARTERIES

184. 263. 185. 271. 173. 230. 222. 215. 233. 212. 221. 239. 168.
231. 221. 131. 211. 232. 313. 240. 176. 210. 251. 175. 185. 184.
198. 198. 208. 284. 231. 171. 258. 164. 230. 197. 216. 230. 265.
197. 230. 233. 250. 243. 175. 200. 240. 185. 213. 180. 208. 386.
236. 230. 188. 200. 212. 193. 230. 169. 181. 189. 180. 297. 232.
150. 239. 178. 242. 323. 168. 197. 417. 172. 240. 191. 217. 208.
220. 191. 119. 171. 179. 208. 180. 254. 191. 176. 283. 253. 220.
268. 248. 245. 171. 239. 196. 247. 219. 159. 200. 233. 232. 189.
237. 319. 171. 194. 244. 236. 260. 254. 250. 196. 298. 306. 175.
251. 255. 285. 184. 228. 171. 229. 195. 214. 221. 204. 276. 165.
211. 264. 245. 227. 197. 196. 193. 211. 185. 157. 224. 209. 223.
278. 251. 140. 197. 172. 174. 192. 221. 283. 178. 185. 181. 191.
185. 206. 210. 226. 219. 215. 228. 245. 186. 242. 201. 239. 179.
218. 279. 234. 264. 237. 162. 245. 191. 207. 248. 139. 246. 247.
193. 332. 194. 195. 243. 271. 197. 242. 175. 138. 244. 206. 191.
223. 172. 190. 144. 194. 105. 201. 193. 262. 211. 178. 331. 235.
267. 227. 243. 261. 185. 171. 222. 231. 258. 211. 249. 209. 177.
165. 299. 274. 219. 233. 220. 348. 194. 230. 250. 173. 260. 258.
131. 168. 208. 287. 308. 227. 168. 178. 164. 151. 165. 249. 258.
194. 140. 187. 171. 221. 294. 167. 208. 208. 185. 159. 222. 266.
217. 249. 218. 245. 242. 262. 169. 204. 184. 206. 198. 242. 189.
260. 199. 207. 206. 210. 229. 232. 267. 228. 187. 304. 140. 209.
198. 270. 188. 160. 218. 257. 259. 139. 213. 178. 172. 198. 222.
238. 273. 131. 233. 269. 170. 149. 194. 142. 218. 194. 252. 184.
203. 239. 232. 225. 280. 185. 163. 216.

concentrations, but that a significant number corresponding to contours in the region of M2 could be classified as diseased. An interpretation was then given to explain the effect of triglyceride concentrations in ascertaining the presence of coronary artery disease, "over and above that implied by the co-existing levels of plasma cholesterol alone." In their analysis, Scott, et al., chose to delete one outlier from each of the TRG data sets based on normality

Table 20b. Listing of Coronary Data - Diseased

PLASMA TRIGLYCERIDES - DISEASE IN AT LEAST 1 OF 3 CORONARY ARTERIES

145.	142.	115.	128.	56.	304.	151.	168.	340.	171.	140.	97.	131.
145.	432.	137.	124.	258.	256.	221.	166.	92.	189.	148.	256.	222.
149.	333.	112.	245.	181.	165.	210.	76.	492.	87.	112.	90.	156.
158.	146.	142.	118.	50.	489.	68.	196.	116.	130.	80.	220.	162.
152.	162.	220.	101.	130.	188.	158.	112.	104.	84.	202.	232.	328.
426.	154.	100.	144.	195.	208.	291.	198.	140.	441.	115.	327.	262.
75.	115.	84.	170.	126.	149.	102.	153.	136.	217.	424.	222.	172.
154.	312.	120.	108.	92.	141.	137.	454.	125.	152.	127.	131.	135.
400.	418.	78.	183.	108.	148.	144.	170.	161.	130.	143.	408.	153.
117.	271.	930.	255.	142.	120.	242.	137.	223.	268.	150.	199.	121.
91.	259.	446.	146.	265.	103.	170.	122.	120.	59.	124.	82.	80.
152.	152.	164.	101.	106.	117.	101.	179.	199.	109.	168.	119.	233.
130.	133.	217.	72.	267.	325.	130.	257.	273.	85.	297.	137.	126.
123.	317.	135.	269.	88.	91.	166.	90.	316.	142.	173.	87.	91.
290.	250.	116.	363.	112.	89.	347.	179.	246.	91.	177.	201.	149.
154.	207.	120.	125.	125.	36.	92.	259.	88.	304.	84.	134.	144.
199.	202.	126.	174.	100.	90.	229.	161.	328.	306.	256.	89.	133.
151.	93.	323.	163.	101.	153.	154.	400.	137.	160.	300.	127.	151.
61.	91.	77.	209.	260.	172.	126.	101.	80.	73.	155.	146.	145.
196.	99.	390.	135.	156.	135.	80.	201.	148.	231.	82.	108.	164.
227.	200.	207.	322.	180.	169.	158.	84.	182.	148.	124.	248.	176.
98.	153.	150.	107.	95.	296.	583.	192.	149.	115.	149.	102.	376.
105.	110.	148.	125.	96.	402.	240.	54.	261.	125.	146.	103.	348.
156.	146.	96.	141.	84.	284.	237.	272.	111.	567.	278.	233.	184.
170.	38.	161.	240.	218.	110.	156.	101.					

considerations (number of standard deviations from the mean). The corresponding CHL values then must be eliminated in the bivariate analysis.

For this data set (data set H), a BISAM analysis was carried out for the normal and diseased groups. Scatter plots for these two groups appear in Figures 30 and 31, and several density estimates were computed to obtain the shapes depicted in Figures 32 through 36. The

Table 21. Summary of BISAM Output for Coronary Data - Normal

Coefficients for Bivariate Dependence Density (1 outlier omitted):

NU1	NU2	REAL(COF)	IMAG(COF)
0	-1	-0.1352	-0.0478
0	1	-0.1352	0.0478
-1	0	-0.1061	0.0630
1	0	-0.1061	-0.0630
-1	1	0.0315	-0.0563
1	-1	0.0315	0.0563
-1	-1	-0.1071	0.0676
1	1	-0.1071	-0.0676

Integrating Factor = 1.015

CHL Mean	195.14	TRG Mean	140.33
CHL Median	195.14	TRG Median	120.00
CHL Variance	1308.38	TRG Variance	5504.91
CHL AR Order	0	TRG AR Order	2
The following are computed with one outlier omitted:			
Pearson r	0.188	H(d-tilda)	-0.026
Spearman Rho	0.243	H(d8)	-0.099
Kendall Tau-A	0.166	H(d24)	-0.116
Kendall Tau-B	0.167	H(d48)	-0.681

univariate density-quantile plots appear in Figures 37 through 40. One rejects normality for both TRG data sets. For this analysis, both normal and diseased groups are classified as bimodal with modes for the normal group at (190,97) and (206,146), and modes for the diseased group at (187,120) and (221,145). The two modes for the diseased group support the results of Scott, et al., but the normal results are contradictory. This analysis was performed for the complete data, with bimodal TRG densities.

Upon eliminating the outliers suggested above, it was discovered

Table 22. Summary of BISAM Output for Coronary Data - Diseased

Coefficients for Bivariate Dependence Density (1 outlier omitted):

NU1	NU2	REAL(COF)	IMAG(COF)
0	-1	-0.0489	0.0546
0	1	-0.0489	-0.0546
-1	0	-0.0333	0.0221
1	0	-0.0333	-0.0221
-1	1	0.1024	-0.0579
1	-1	0.1024	0.0579
-1	-1	-0.0148	0.0472
1	1	-0.0148	-0.0472
0	-2	0.0441	-0.0498
0	2	0.0441	0.0498
-2	0	-0.0198	-0.0026
2	0	-0.0198	0.0026
-1	2	0.0831	-0.0029
1	-2	0.0831	0.0029
-2	1	-0.0049	0.0821
2	-1	-0.0049	-0.0821
-2	-1	-0.0377	-0.0252
2	1	-0.0377	0.0252
-1	-2	-0.0411	-0.0269
1	2	-0.0411	0.0269
-2	2	0.0560	-0.0118
2	-2	0.0560	0.0118
-2	-2	0.0626	0.0230
2	2	0.0626	-0.0230

Integrating Factor = 1.009

CHL Mean	216.19	TRG Mean	179.35
CHL Median	212.50	TRG Median	150.00
CHL Variance	1850.04	TRG Variance	10372.6
CHL AR Order	0	TRG AR Order	1
The following are computed with one outlier omitted:			
Pearson r	0.210	H(d-tilda)	-0.300
Spearman Rho	0.270	H(d8)	-0.078
Kendall Tau-A	0.183	H(d24)	-0.105
Kendall Tau-B	0.184	H(d48)	-0.158

that the results for the diseased group were fairly stable, but for

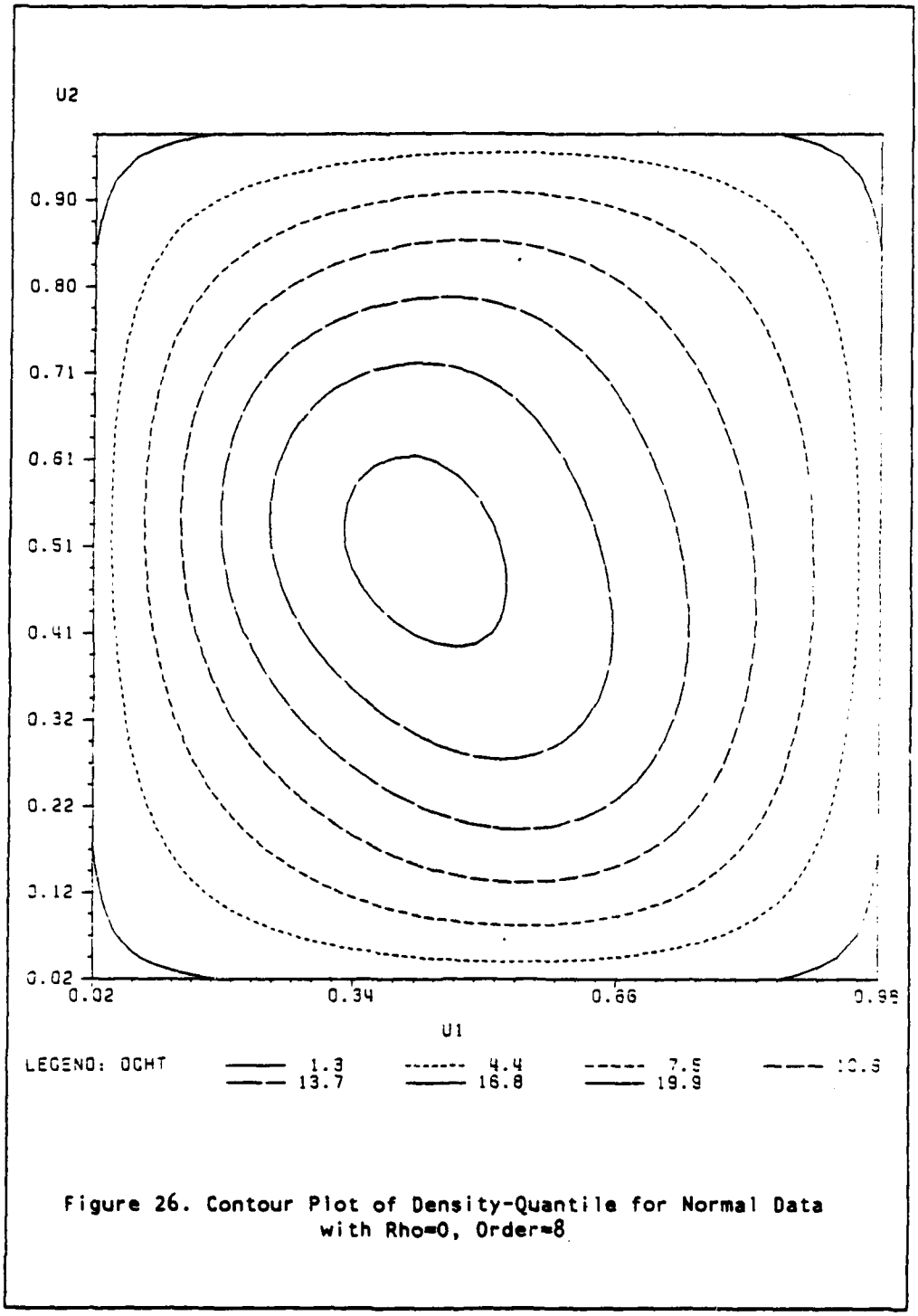


Figure 26. Contour Plot of Density-Quantile for Normal Data with Rho=0, Order=8

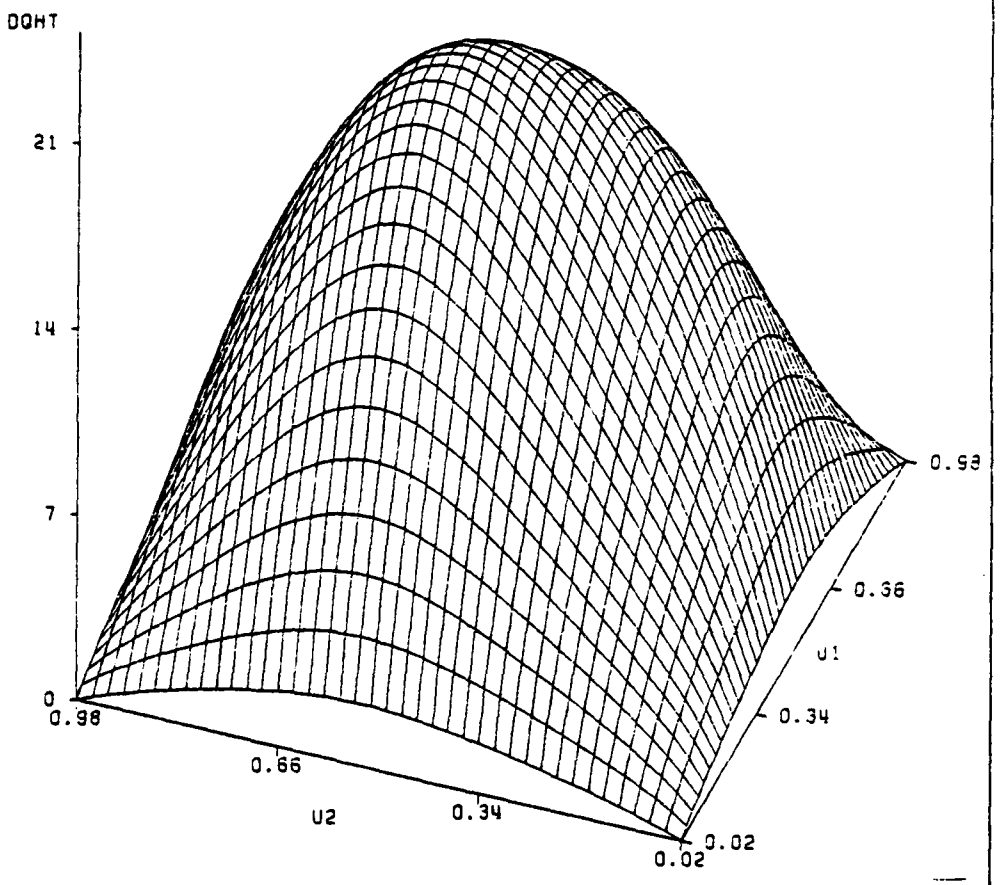


Figure 27. Three Dimensional Plot of Density-Quantile for Normal Data with Rho=0, Order=8

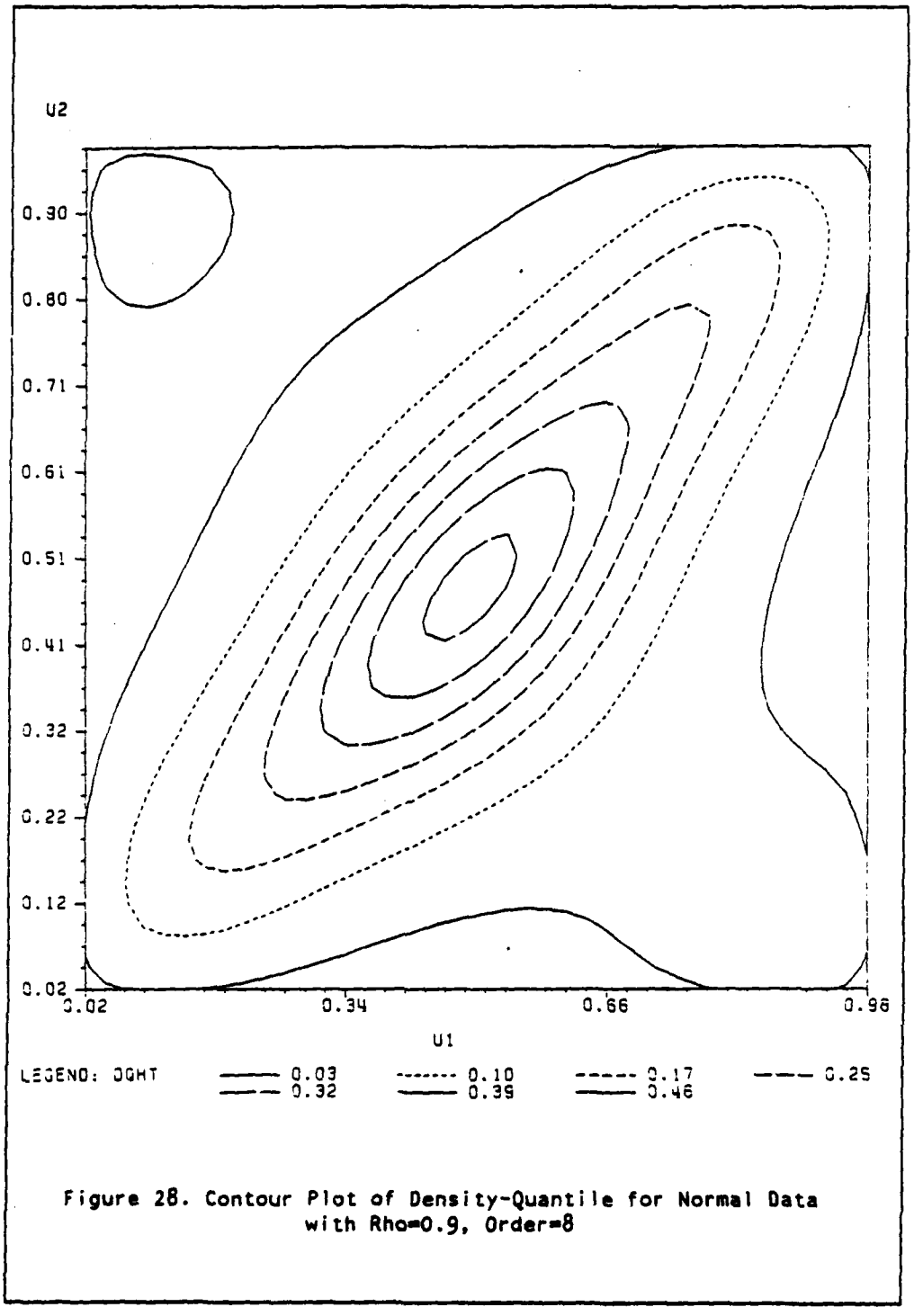


Figure 28. Contour Plot of Density-Quantile for Normal Data with Rho=0.9, Order=8

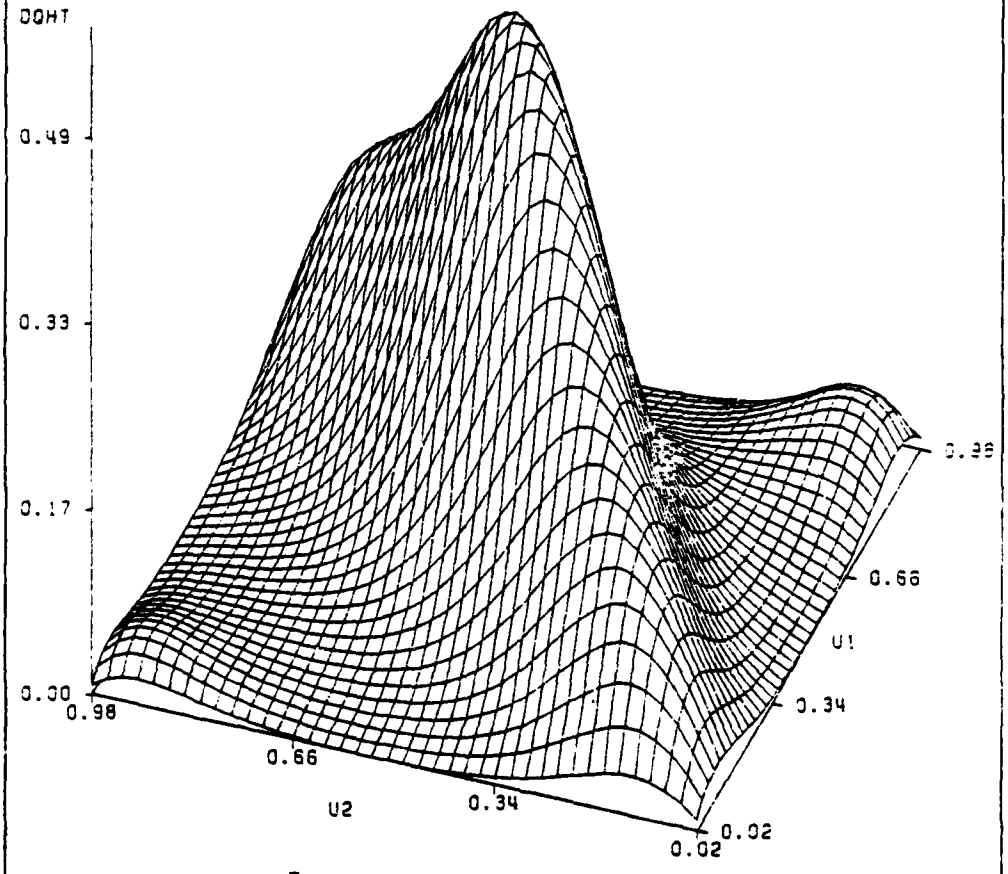


Figure 29. Three Dimensional Plot of Density-Quantile for Normal Data with Rho=0.9, Order=8

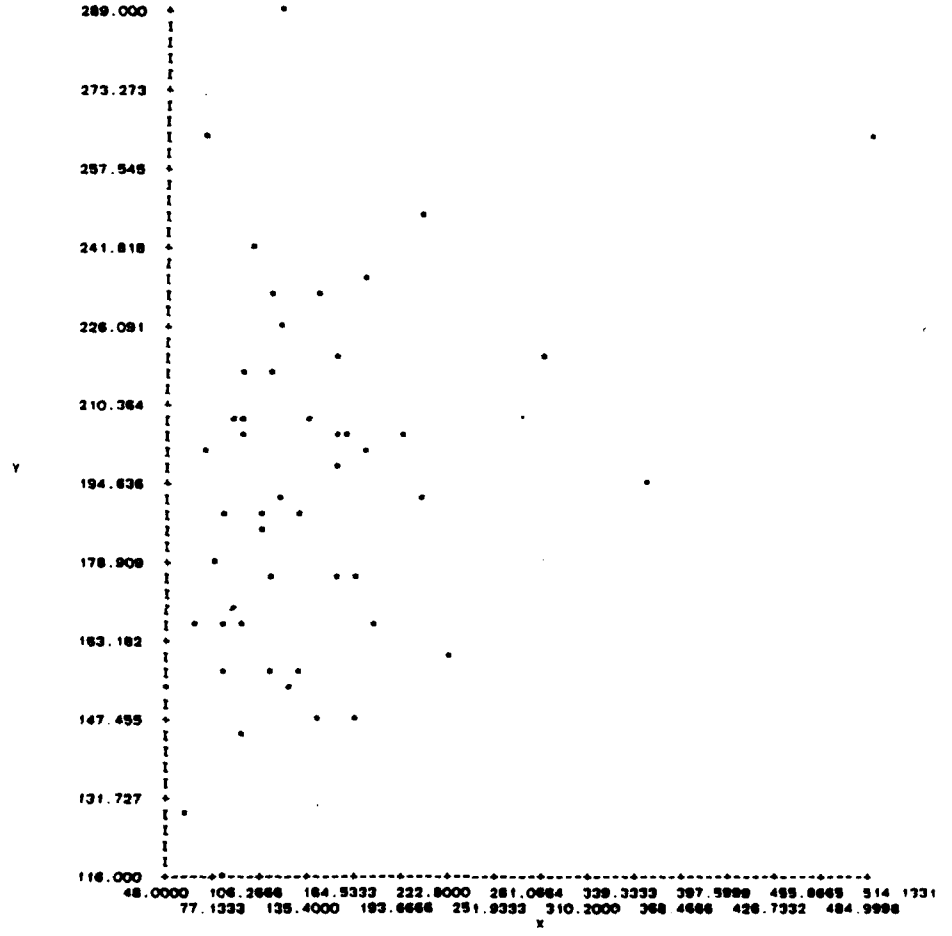


Figure 30. Scatter Plot of Coronary Data, Normal

the normal group, a unimodal shape was obtained corresponding to that of Scott, et al., For this case, a mode for the normal group occurs at (195,122) and two modes for the diseased group occur at (188,120) and

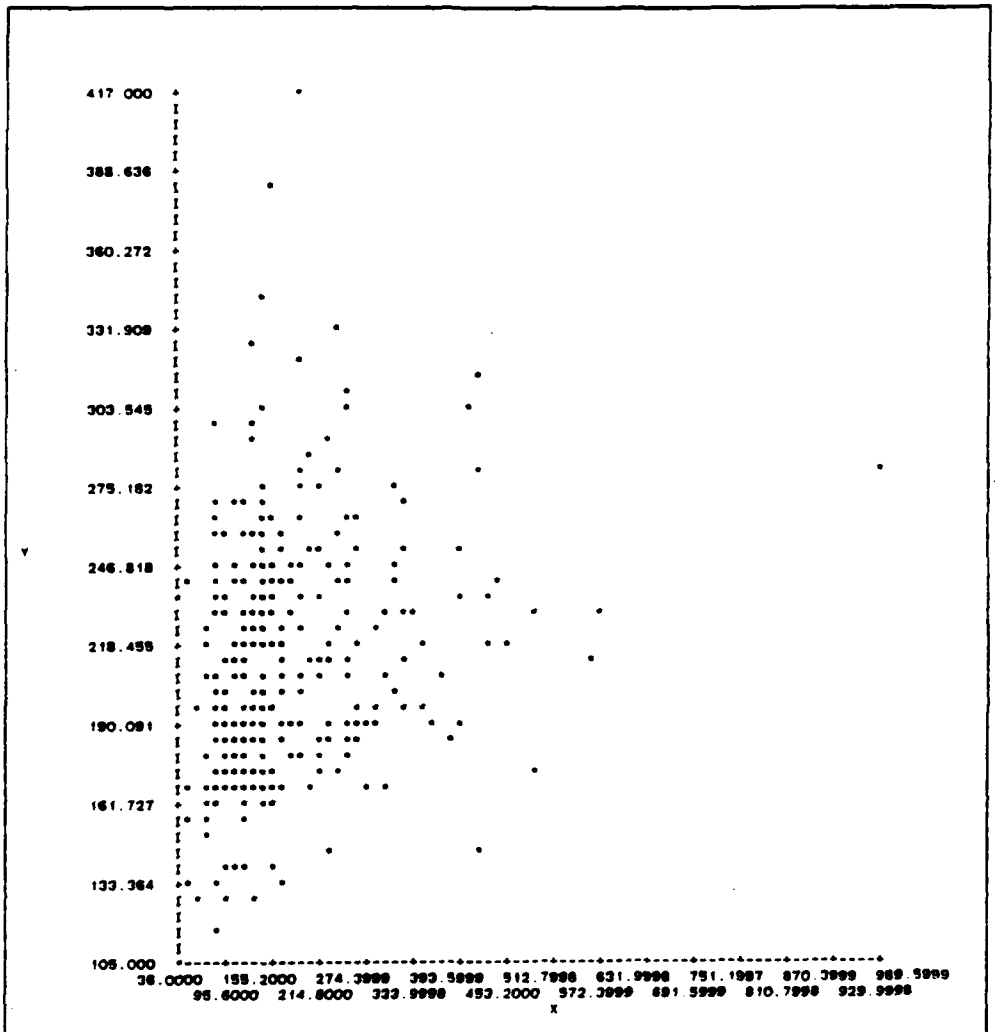


Figure 31. Scatter Plot of Coronary Data, Diseased

(221,145). These results seem to confirm those of Scott, et al., Elimination of the two outliers served to produce unimodal univariate densities in all cases. Figures 41 through 45 illustrate the

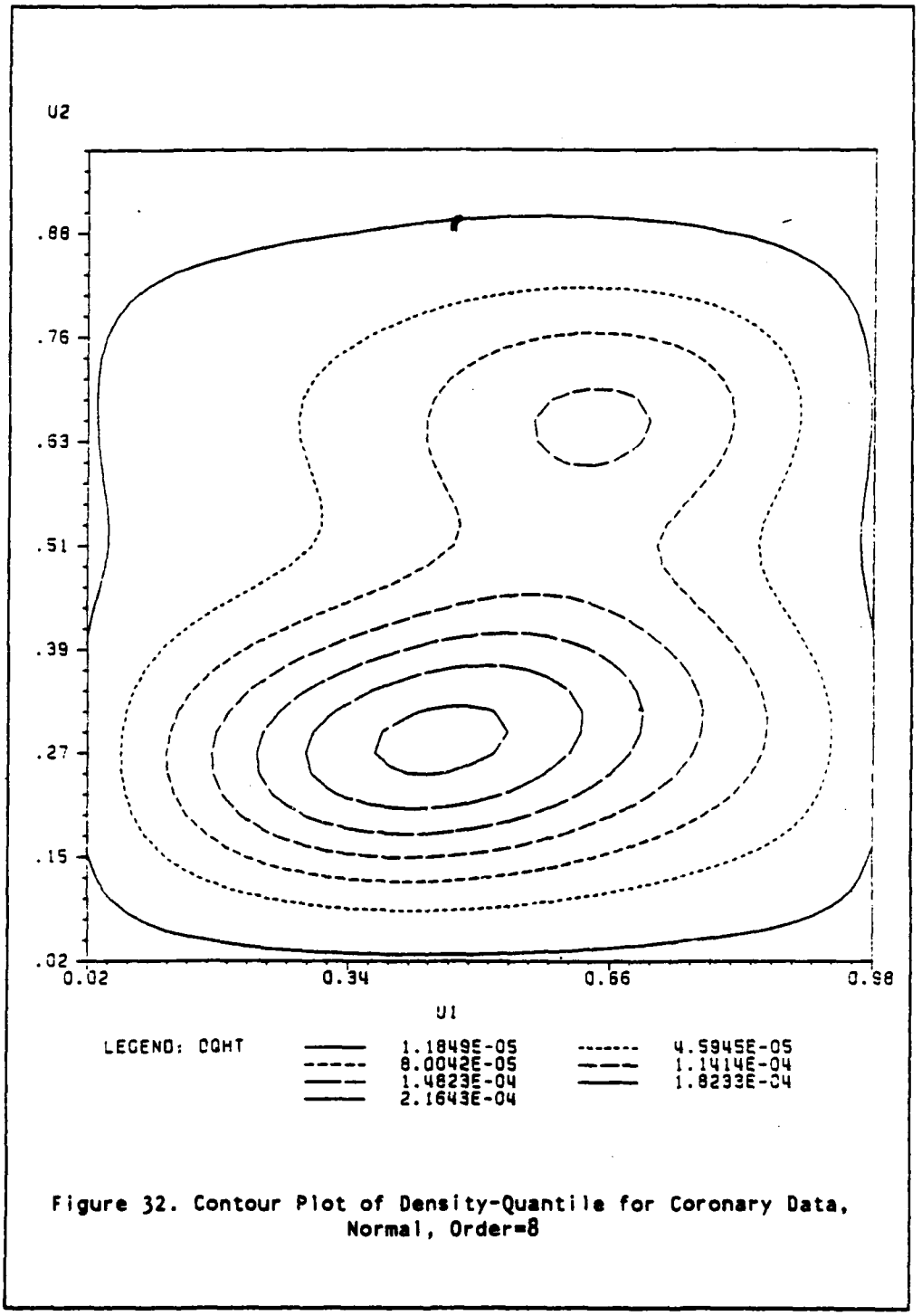


Figure 32. Contour Plot of Density-Quantile for Coronary Data, Normal, Order=8

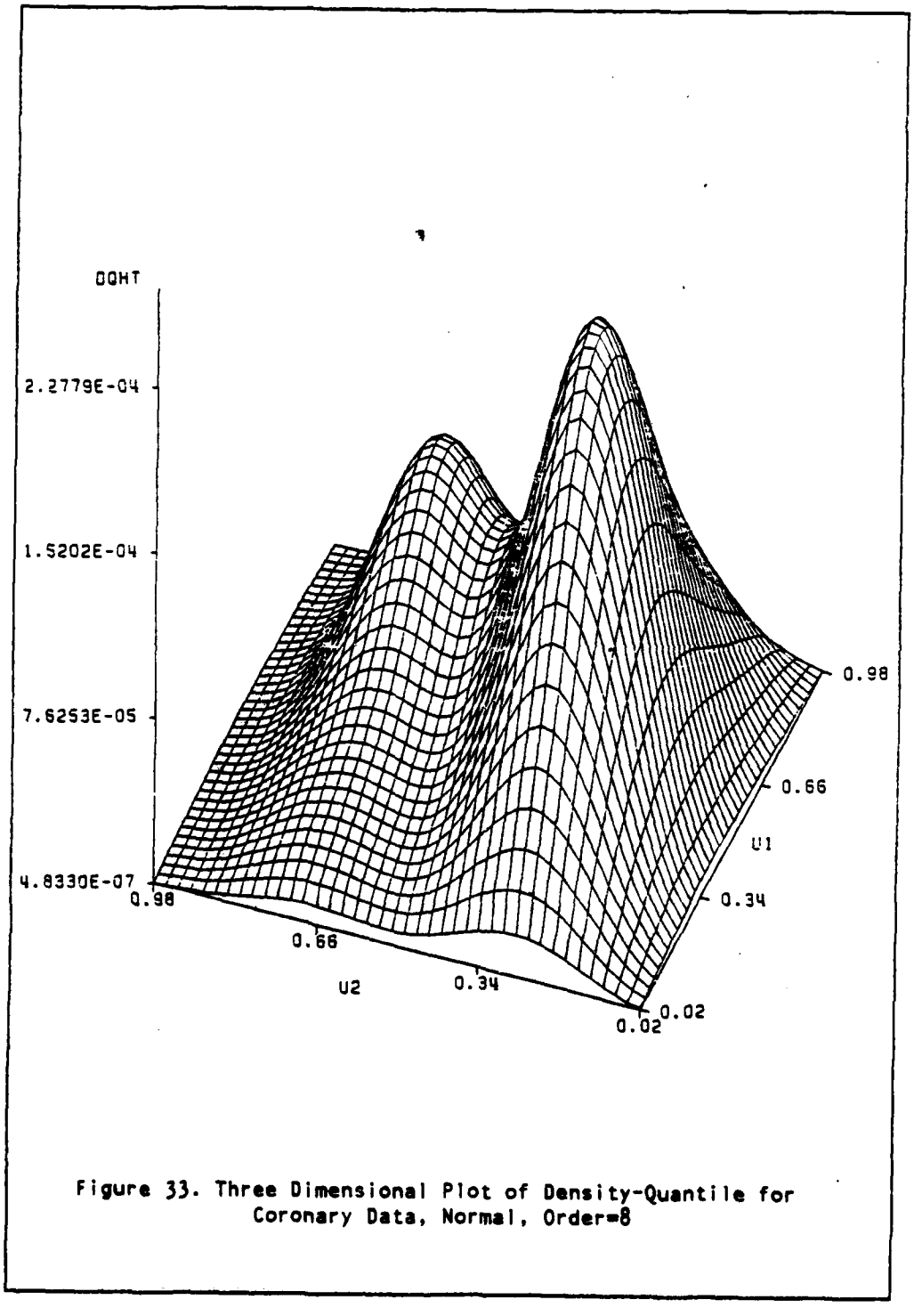


Figure 33. Three Dimensional Plot of Density-Quantile for Coronary Data, Normal, Order=8

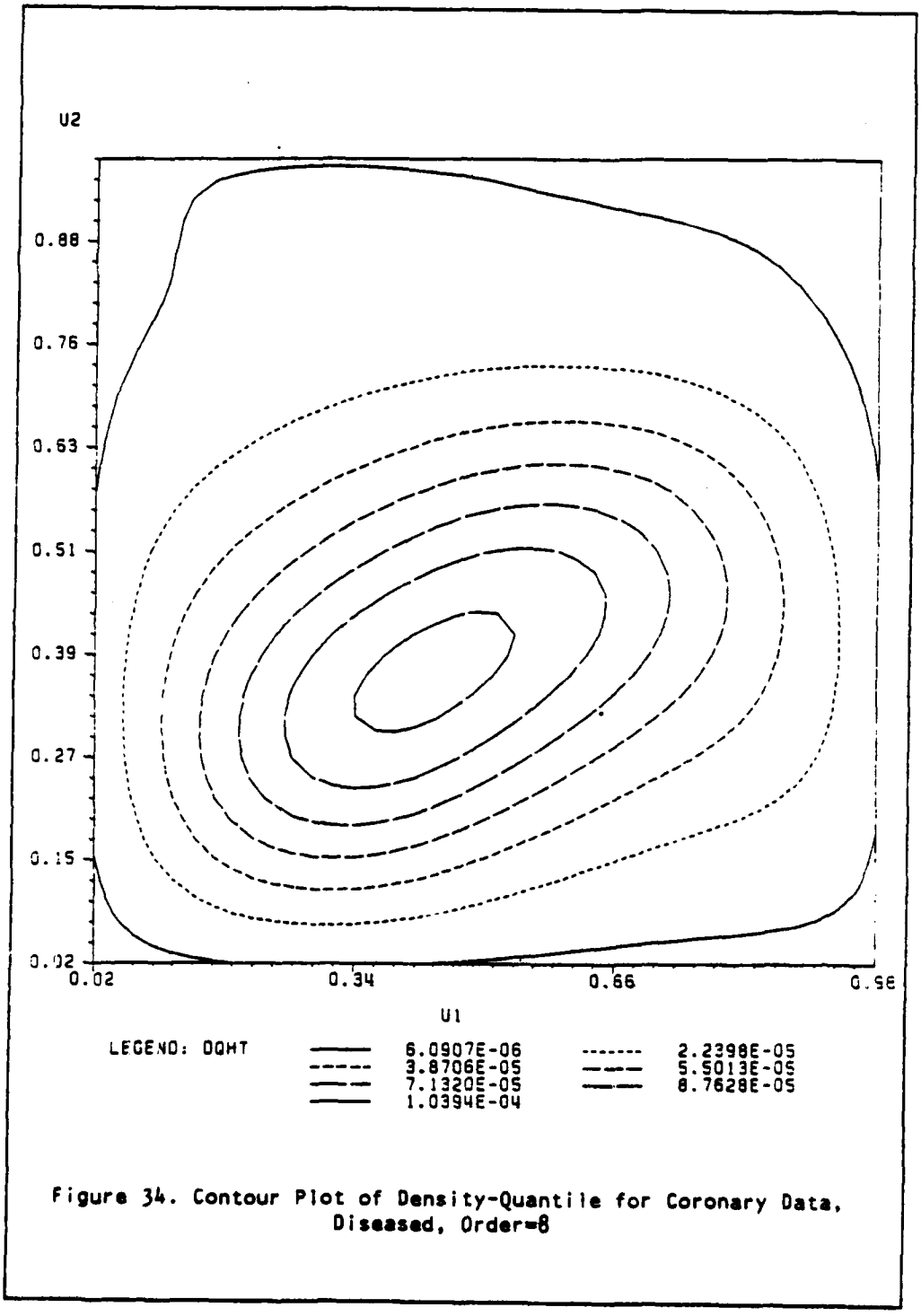


Figure 34. Contour Plot of Density-Quantile for Coronary Data, Diseased, Order=8

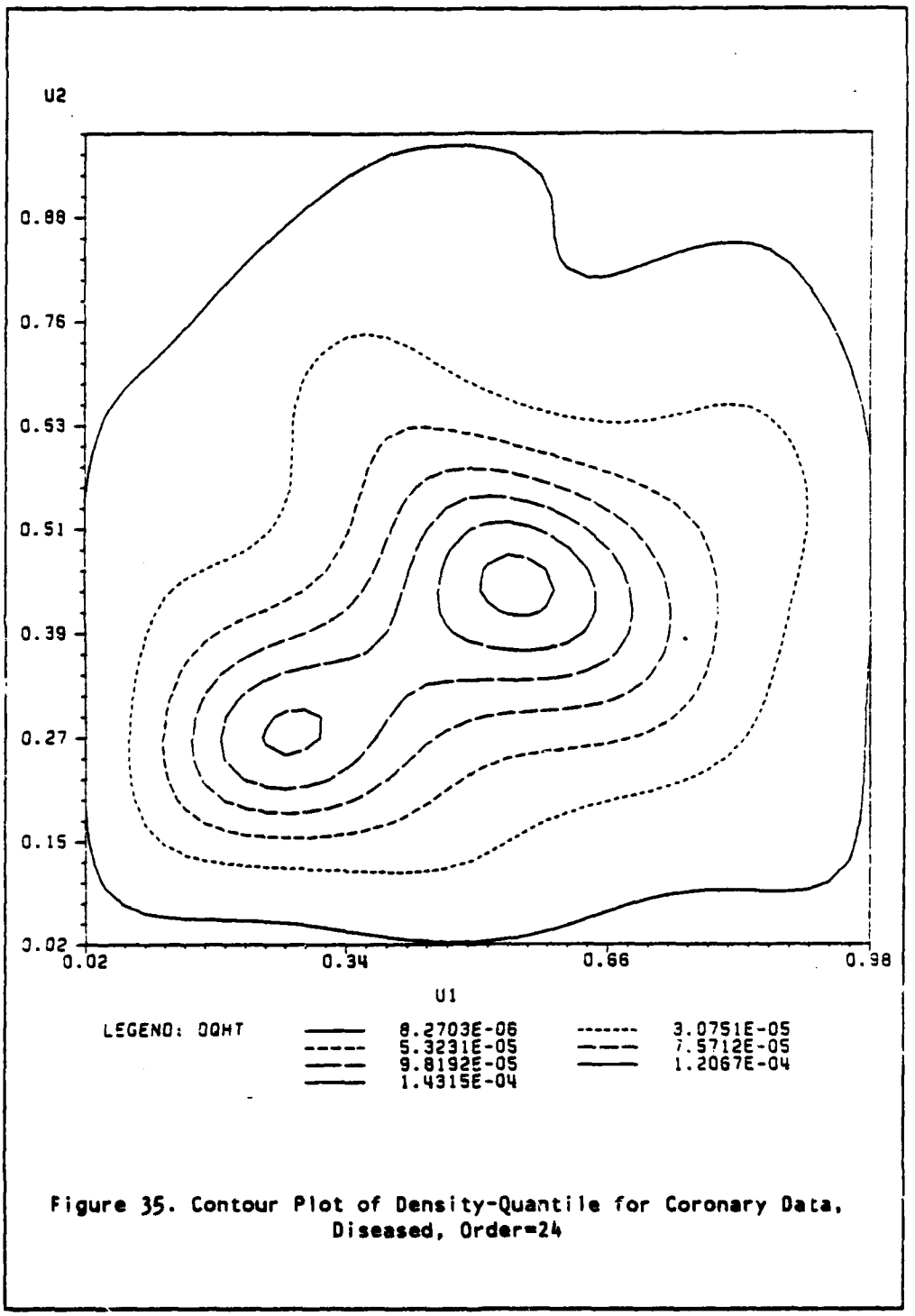


Figure 35. Contour Plot of Density-Quantile for Coronary Data, Diseased, Order=24

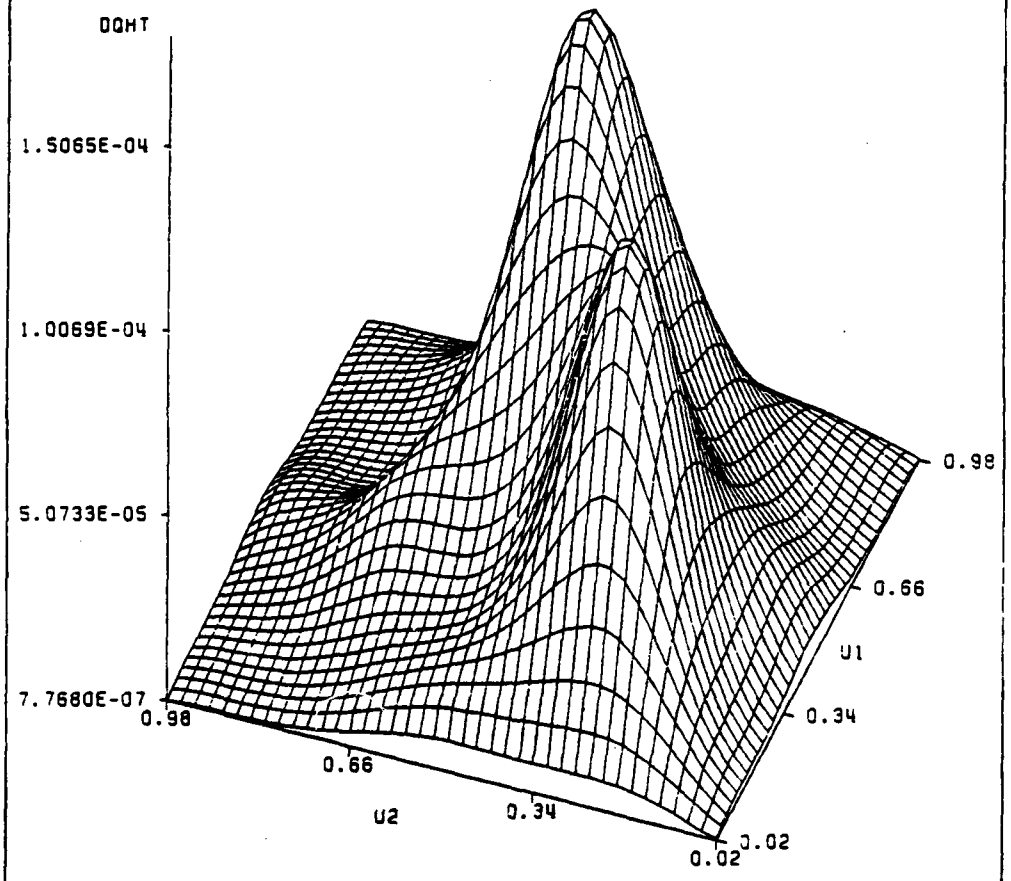


Figure 36. Three Dimensional Plot of Density-Quantile for Coronary Data, Diseased, Order=24

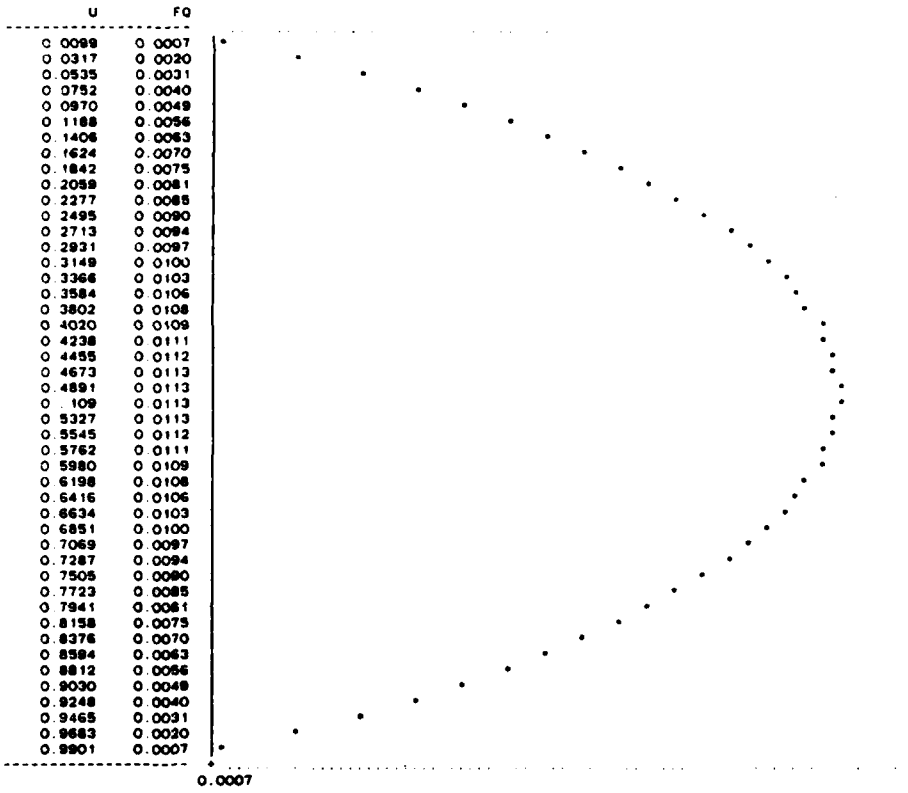


Figure 37. Univariate Density-Quantile Plot for Variable CHL, Normal

bivariate shapes obtained when one outlier is removed.

This analysis illustrates some important points. Bimodal univariate densities will usually produce multimodal bivariate densities, but it is possible for the dependence structure for two unimodal univariate densities to induce multimodal bivariate

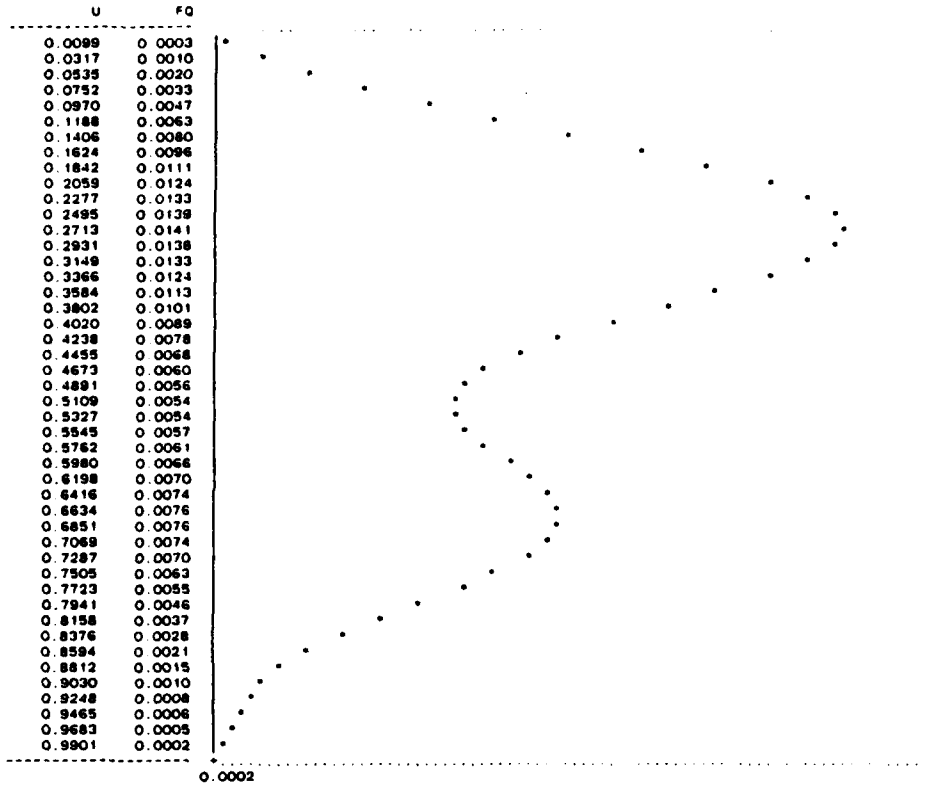


Figure 38. Univariate Density-Quantile Plot for Variable TRG, Normal

densities. The ability of the kernel method and the minimum information method to discover these modes is crucial to a study designed to determine if more than one population is represented by a data set. Furthermore, this analysis illustrates the importance of univariate density estimation in such "bump-hunting" problems. When

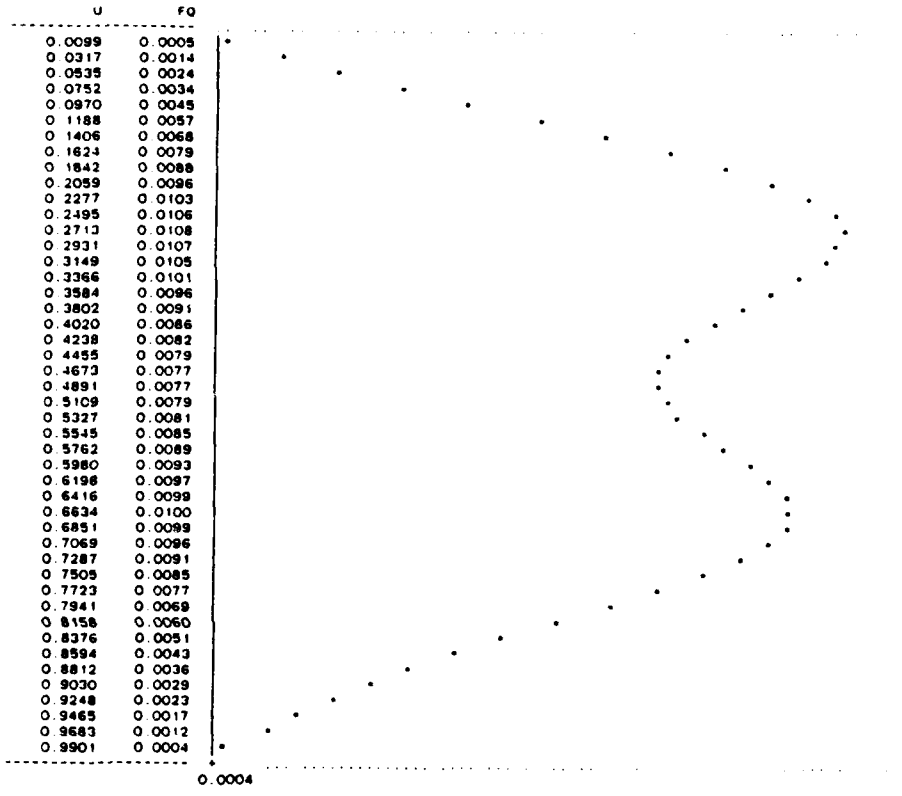


Figure 39. Univariate Density-Quantile Plot for Variable CHL, Diseased

the correlation is small, the univariate densities may dominate the shaping of the bivariate densities. For the coronary data, the correlations range from 0.2 to 0.3 indicating small but statistically significant correlation between the variables CHL and TRG. Consequently, the univariate density-quantile functions have a

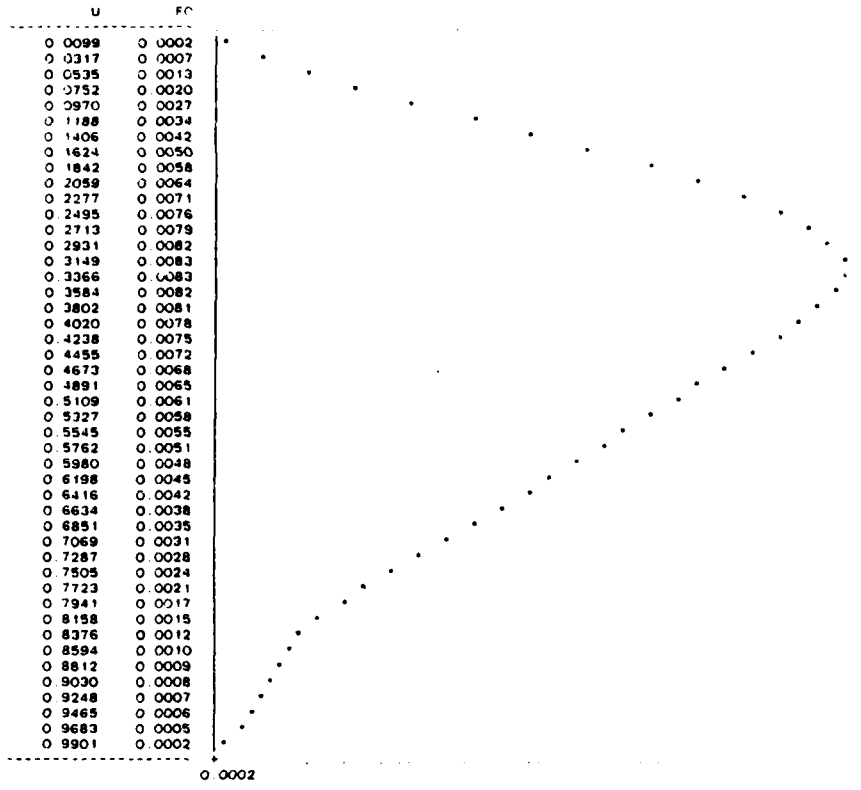


Figure 40. Univariate Density-Quantile Plot for Variable TRG, Diseased

pronounced effect on the bivariate density-quantile, especially for the normal group. When the outlier is omitted from the normal group, the shape of the TRG density changes from bimodal to unimodal with a corresponding change in the bivariate density-quantile. Clearly a univariate analysis is crucial to bivariate data modeling, a fact that

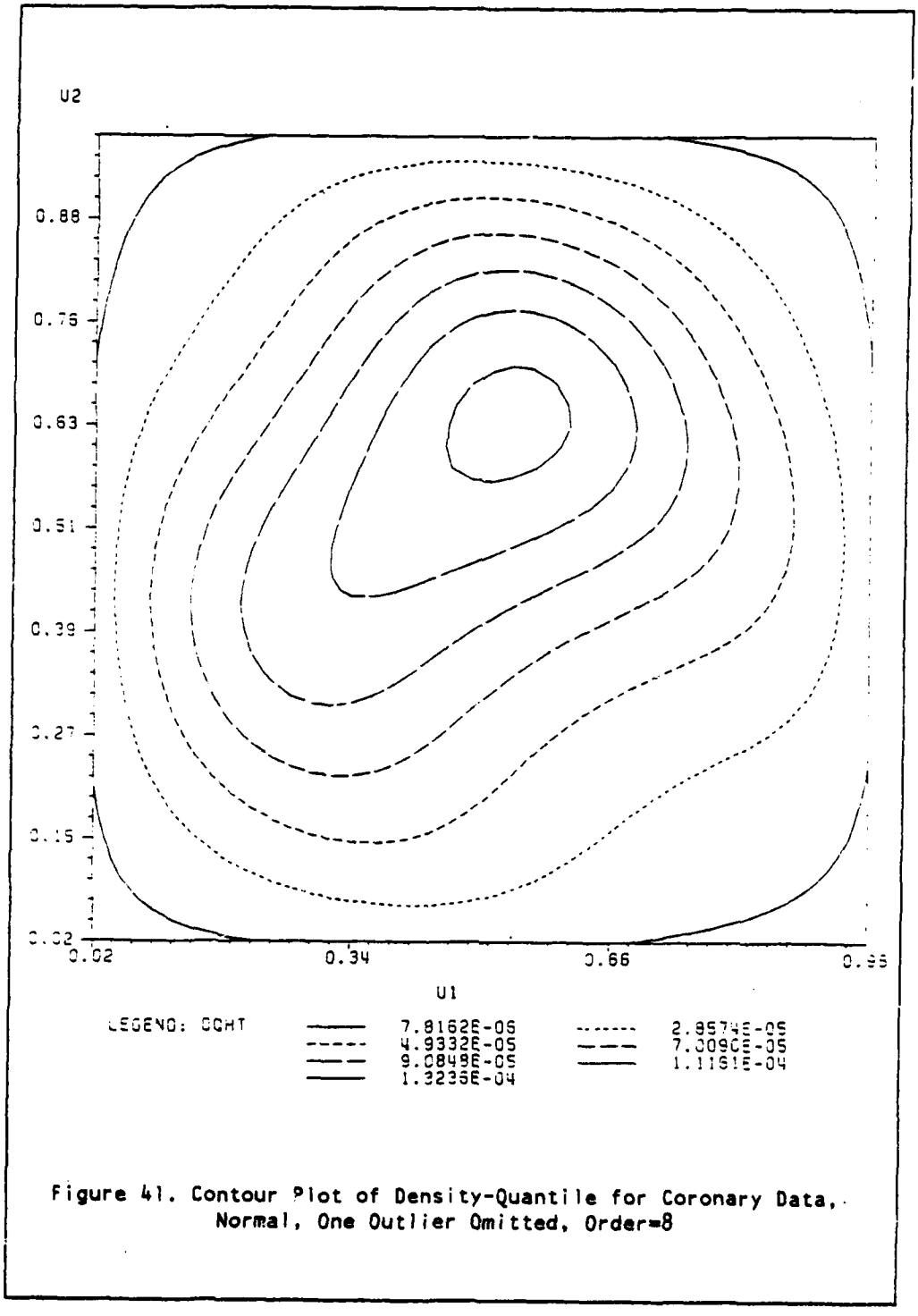


Figure 41. Contour Plot of Density-Quantile for Coronary Data, Normal, One Outlier Omitted, Order=8

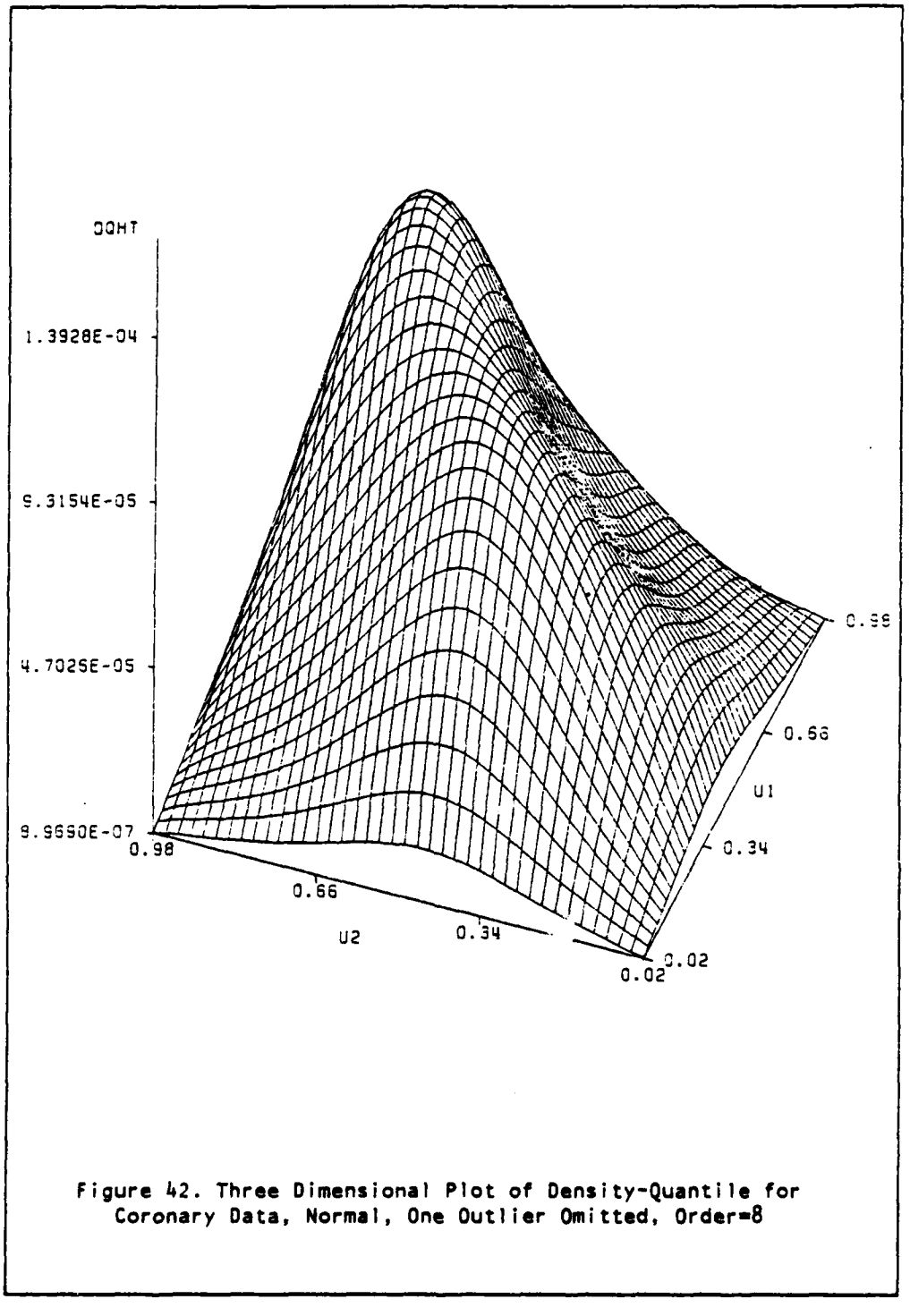


Figure 42. Three Dimensional Plot of Density-Quantile for Coronary Data, Normal, One Outlier Omitted, Order=8

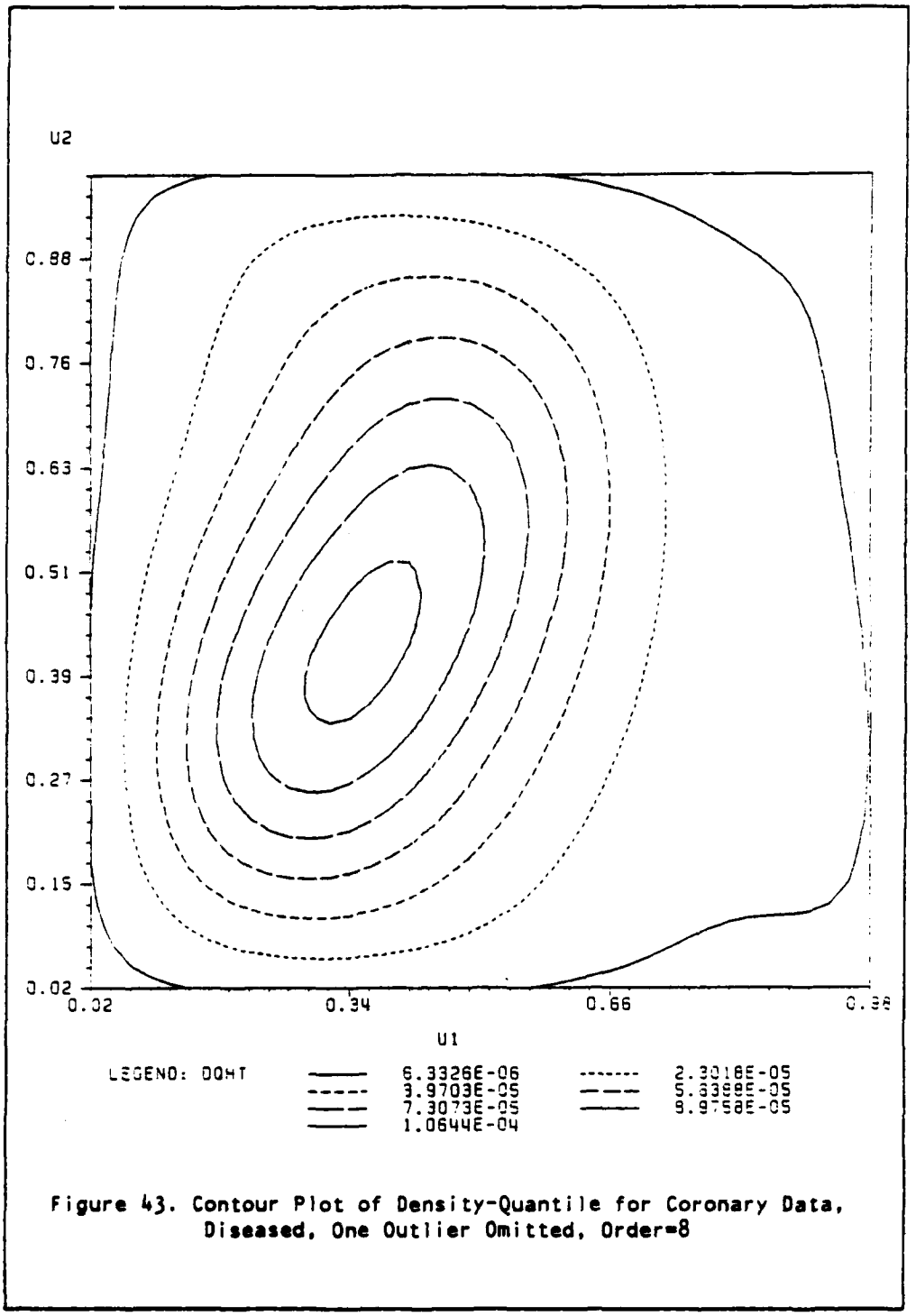


Figure 43. Contour Plot of Density-Quantile for Coronary Data, Diseased, One Outlier Omitted, Order=8

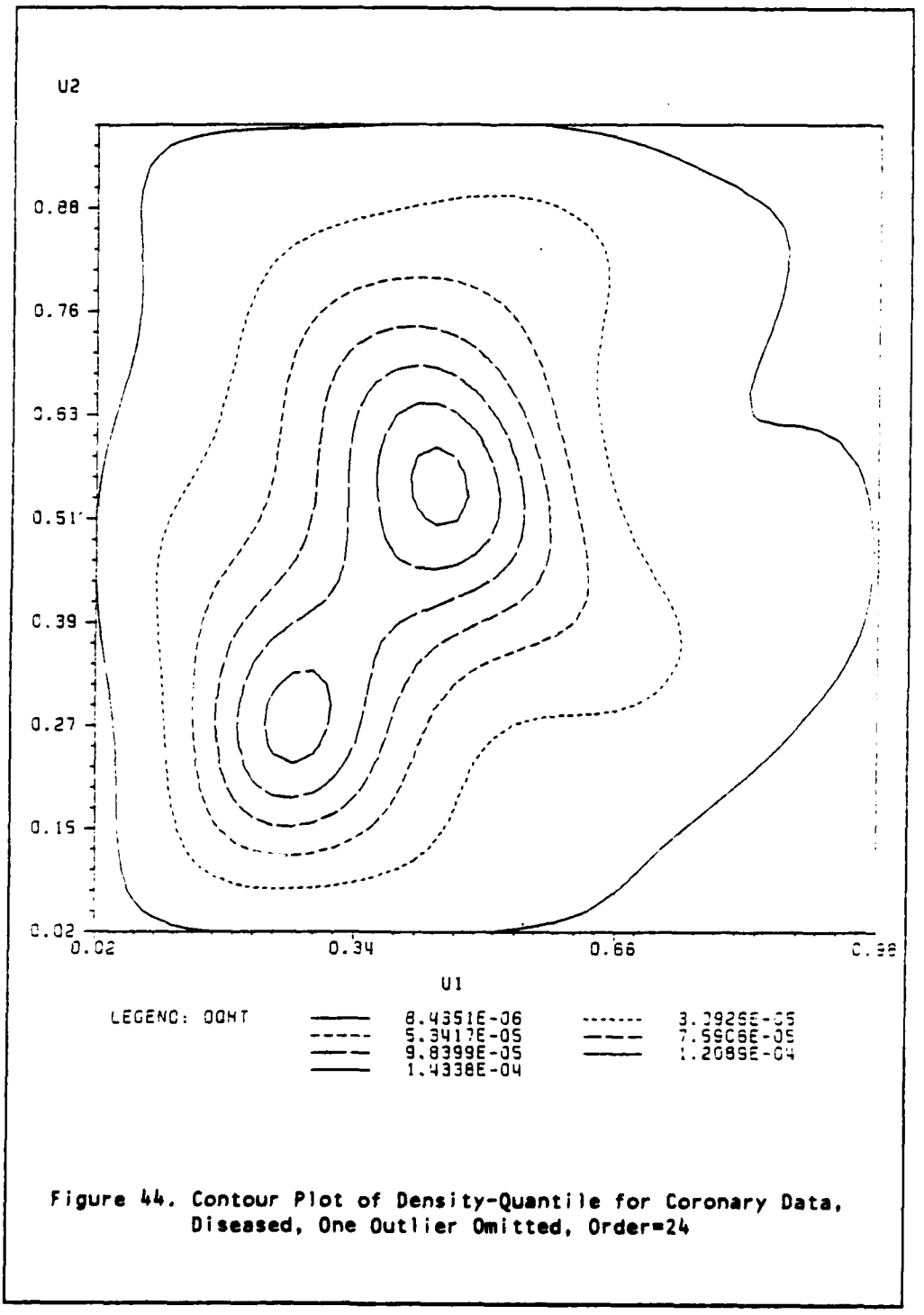


Figure 44. Contour Plot of Density-Quantile for Coronary Data, Diseased, One Outlier Omitted, Order=24

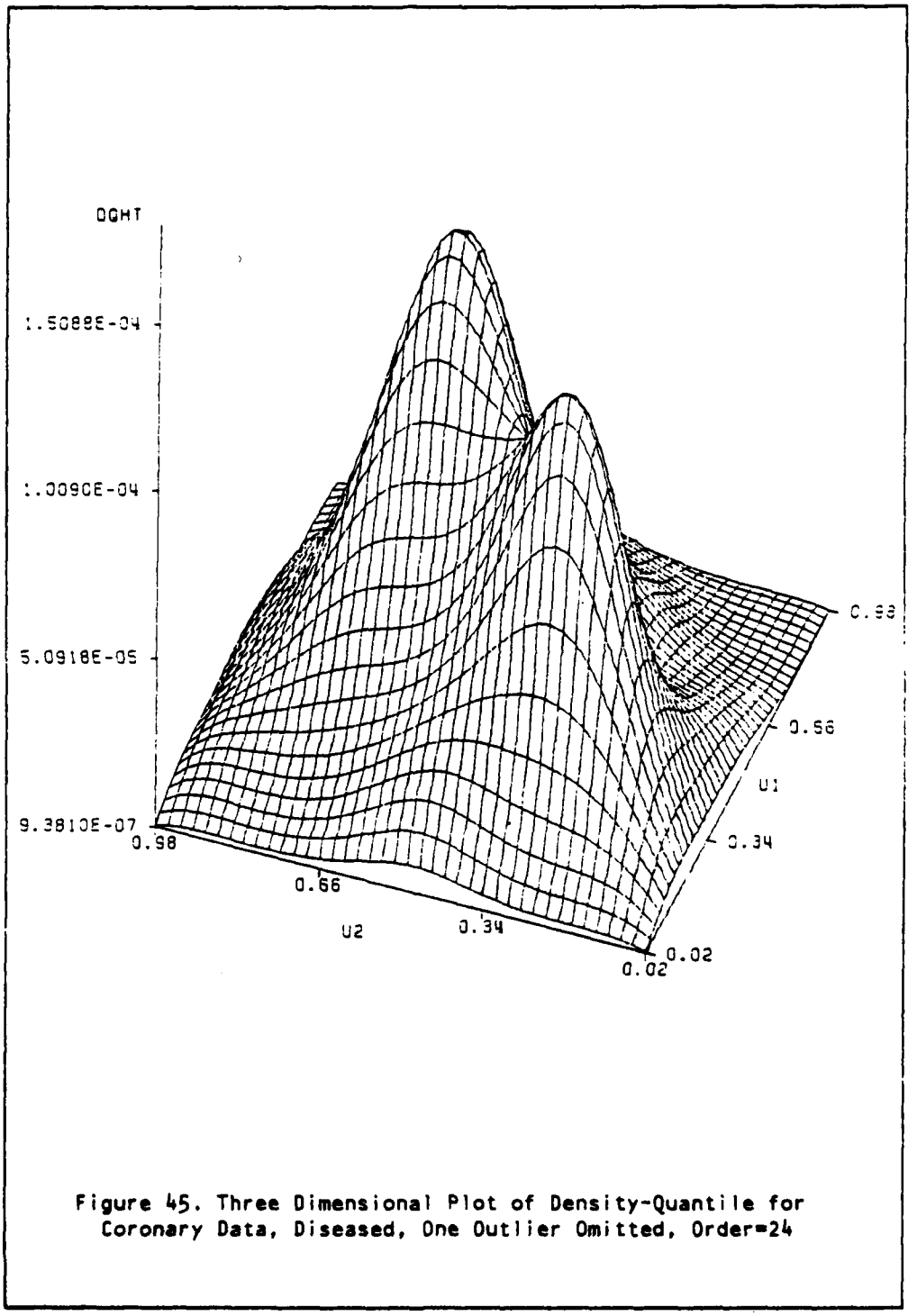


Figure 45. Three Dimensional Plot of Density-Quantile for Coronary Data, Diseased, One Outlier Omitted, Order=24

may be hidden by the kernel approach.

These examples serve to illustrate the competitiveness of minimum information bivariate density estimation and point out some fundamental characteristics of a bivariate data modeling approach. For any exploratory analysis one is especially concerned with obtaining all of the possible shapes that can effectively model a data set. The minimum information approach provides such a multitude of shapes with a minimum amount of expurgatorial effort that occurs in such approaches as the kernel method. However, the approach would be aided by an objective procedure to choose among the orders of approximation. This is left as a subject for further study with the failures noted herein serving as a catapult for future research.

7. CONCLUSION

7.1 Concluding Remarks

In the dissertation we have shown the utility of the function approximation approach to density estimation. The representation of an unknown density as a truncated parametric model that is nonparametric in its assumptions for the data provides a useful tool for an exploratory analysis of a data set. The convergence of infinite series representations is usually very rapid so that truncated series provide good approximations for functions of interest. The statistical problem involves estimating the parameters of the expansion in a stochastic setting. Unfortunately, an exact stochastic model to aid in estimation is not always available, and asymptotic results may only be meaningful for very large samples. Consequently, the properties of estimates are difficult to investigate in such a general setting and hence restrict the applicability of such approaches to inferential statistics. Simulation studies and experience in the application of expansion techniques, however, support their use in the absence of exact theory.

Heuristic motivations are given by considering finite exponential models and their relationship to Fourier series in Hilbert space. Analogies to classical regression and time series analysis also provide heuristic support to this approach. The development of information functionals then serve to promote the regression analogy by suggesting least squares estimation of parameters. Information

theory plays a dual role in providing criterion functionals and parameters that represent important population characteristics. One suspects that the observation of Berkson (1980) concerning the selection of estimation criteria have a more general interpretation in terms of information functionals. The value of information theory as applied to statistics has yet to be fully realized, but works such as this one should motivate further research into the area.

In this work we have presented some fundamental mathematical and statistical concepts useful to the study of bivariate data modeling. The analysis of a variety of nonparametric density estimation procedures suggested problems to be overcome and motivated the development of a new technique that competes with existing procedures and extends easily to the bivariate case. Classical inference procedures for testing for independence were explored with suggestions made for the development of entropy statistics that measure association between two random variables. An information parameter was derived that represented a measure of association, but satisfactory estimators of the parameter were not obtained. A bivariate data modeling approach was investigated and promising results obtained for the problem of density estimation and "bump-hunting" in a multivariate setting. The development of univariate and bivariate density estimation programs provided interesting comparisons and permitted the exploratory analysis of a variety of data sets. The program BISAM permits application of the data modeling results obtained and provides a useful computing tool to the applied statistician. The BISAM program forms a computing triad

with ONESAM (Parzen and Anderson, 1980) and TWOSAM (Prihoda, 1981) that may serve as a valuable tool to the data analyst. The nonparametric and robust procedures available from these programs are difficult to obtain from one source and overcome the weaknesses of most statistical packages in the areas of data analysis and density estimation. This contribution should motivate the application of nonparametric procedures to statistical problems and should support the continuing research into multivariate nonparametric methods.

7.2 Problems for Further Study

The entropy estimates of section 4.5 were disappointing in comparison with existing measures of association, but it is felt that better estimates of parameter H(d) will make this an important approach in testing for independence. Furthermore, entropy diagnostics should be valuable tools to many areas of bivariate data modeling. For example, goodness-of-fit tests are readily suggested by examining the information between a null hypothesis density and a nonparametric density estimate. The utility of information functionals spreads across many areas of application. A characterization problem might be aided by information functionals, with an open research question being whether

$$I(f_{X,Y}; f_X f_Y) = -(1/2)\log(1-\rho^2) \qquad (9.2.1)$$

is a defining characteristic of bivariate normality or whether it

defines a more general class of distributions. If this equation characterized bivariate normality, an entropy based test of bivariate normality could be developed using Pearson's r to estimate the normal entropy and using the estimated dependence density to form an entropy statistic. The difference of these statistics would then be a version of the information divergence discussed in section 2.5.

The use of subjective criteria in model selection should motivate the study of objective techniques for choosing a model over competitors. One is motivated to consider a function of the sample entropy to evaluate the contribution of additional terms in an expansion. A version of maximum entropy analagous to the AIC or CAT criterion functions could then be developed to arrive at an optimal order of approximation.

The distribution theory for entropy statistics remains to be investigated, with motivation provided by recent papers by Stute (1982) and Taniguchi (1980). The asymptotic theory in the univariate case for the minimum information density estimate needs to be resolved, indicating the need for continued research into the distribution theory of non-standard regression models.

Multivariate robustness is also a topic of interest, with papers by Gnanadesikan and Kettenring (1972) and Green (1981) suggesting research questions for the problem of trimming a multivariate data set of outliers. The computer implementation of data trimming procedures is also of interest.

Finally, one is interested in discovering bivariate distributions for which the classical nonparametric measures of association perform

poorly. Further research may then be motivated into the numerical approximation of conditional quantile functions of interest to be used in simulation studies. The problem of simulating data corresponding to existing data as a validity check of an analysis has been approached from a conditional quantile perspective, implying the need for good estimates of the conditional quantile function. Such estimators have yet to be proposed, but one may approach the problem using the function approximation techniques suggested by this work.

BIBLIOGRAPHY

Abramowitz, Milton, and Stegun, Irene A., eds. (1972), Handbook of Mathematical Functions, New York: Dover Publications.

Ahlberg, J.H., Nilson, E.N., and Walsh, J.L. (1967), The Theory of Splines and their Applications, New York: Academic Press.

Anderson, G.D. (1969), "A Comparison of Methods for Estimating a Probability Density Function," unpublished Ph.D. Dissertation, University of Washington.

Bean, Steven J., and Tsokos, Chris P. (1980), "Developments in Nonparametric Density Estimation," International Statistical Review, 48, 267-287.

Berkson, Joseph (1980), "Minimum Chi-square, not Maximum Likelihood!" Annals of Statistics, 8, 457-487.

Bloch, Daniel A., and Gastwirth, Joseph L. (1968), "On a Simple Estimate of the Reciprocal of the Density Function," Annals of Mathematical Statistics, 39, 1083-1085.

Blomqvist, Nils (1950), "On a Measure of Dependence between Two Random Variables," Annals of Mathematical Statistics, 21, 593-600.

Blum, J.R., Kiefer, J., and Rosenblatt, M. (1961), "Distribution Free Tests of Independence Based on the Sample Distribution Function," Annals of Mathematical Statistics, 32, 485-498.

BMDP-79: Biomedical Computer Programs, P-Series (1979), Los angeles: University of California Press.

Bochner, Salomon (1960), Harmonic Analysis and the Theory of Probability, Berkeley, University of California Press.

Brillinger, David R. (1975), Time Series: Data Analysis and Theory, New York: Holt, Rinehart, and Winston.

Cacoullos, Theophilos (1966), "Estimation of a Multivariate Density," Annals of the Institute of Statistical Mathematics, 18, 178-179.

Carmichael, Jean-Pierre (1976), "The Autoregressive Method," unpublished Ph.D. thesis, Statistical Science Division, State University of New York at Buffalo.

Cencov, N.N. (1962), "Evaluation of an Unknown Distribution Density from Observations," Soviet Mathematics, 3, 1559-1562.

Cheng, K.F., and Taylor, R.L. (1980), "On the Uniform Complete

Convergence of Estimates for Multivariate Density Functions and Regression Curves," Annals of the Institute of Statistical Mathematics, 32, 187-199.

Churchill, Ruel V. (1969), Fourier Series and Boundary Value Problems, New York: McGraw-Hill.

Conover, W. (1971), Practical Nonparametric Statistics, New York: John Wiley and Sons.

Crain, Bradford (1974), "Estimation of Distributions Using Orthogonal Expansions," Annals of Statistics, 2, 454-463.

Csorgo, M., and Revesz, P. (1981), Strong Approximations in Probability and Statistics, New York: Academic Press.

Davis, Philip J. (1975), Interpolation and Approximation, New York: Dover Publications.

Devroye, L.P., and Wagner, T.J. (1980), "The Strong Uniform Consistency of Kernel Density Estimates," Multivariate Analysis V, P. Krishnaiah, ed., Amsterdam: North Holland.

Dudewicz, Edward J., and Van Der Muelen, Edward C. (1981), "Entropy-based Tests of Uniformity," Journal of the American Statistical Association, 76, 967-974.

Durbin, J. (1973), Distribution Theory for Tests Based on the Sample Distribution Function, Philadelphia: SIAM.

Gibbons, J. (1971), Nonparametric Statistical Inference, New York: McGraw-Hill.

Gnanadesikan, R., and Kettenring, J.R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," Biometrics, 28, 81-124.

Good, I.J., and Gaskins, R.A. (1980), "Density Estimation and Bump-hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data," Journal of the American Statistical Association, 75, 42-73.

Graybill, Franklin A. (1976), Theory and Application of the Linear Model, North Scituate, Massachusetts: Duxbury Press.

Green, P.J. (1981), "Peeling Bivariate Data," Interpreting Multivariate Data, Vic Barnett, ed., New York: John Wiley and Sons.

Hajek, J., and Sidak, Z. (1967), Theory of Rank Tests, New York: Academic Press.

Hewitt, Edwin, and Stromberg, Karl (1965), Real and Abstract Analysis, New York: Springer-Verlag.

Hocking, R.R. (1976), "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32, 1-49.

Hoeffding, Wassily (1948a), "A Class of Statistics with Asymptotically Normal Distributions," Annals of Mathematical Statistics, 19, 293-325.

_____ (1948b), "A Nonparametric Test of Independence," Annals of Mathematical Statistics, 19, 546-557.

Huber, Peter J. (1981), Robust Statistics, New York: John Wiley and Sons.

IMSL (1980), International Mathematical and Statistical Library, Reference Manuel, Vol. II, Houston: IMSL, Inc.

Kendall, M.G. (1938), "A New Measure of Rank Correlation," Biometrika, 30, 81-93.

Kennedy, William J., Jr., and Gentle, James E. (1980), Statistical Computing, New York: Marcel Dekker.

Kimeldorf, George, and Sampson, Allan R. (1975a), "One-Parameter Families of Bivariate Distributions with Fixed Marginals," Communications in Statistics, 4, 293-301.

_____ (1975b), "Uniform Representations of Bivariate Distributions," Communications in Statistics, 4, 617-628.

_____ (1978), "Monotone Dependence," Annals of Statistics 6, 895-903.

Kronmal, R.A., and Tartar, M.E. (1968), "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," Journal of the American Statistical Association, 63, 925-952.

Kullback, Solomon (1978), Information Theory and Statistics, Gloucester, Massachusetts: Peter Smith.

Lanczos, Cornelius (1956), Applied Analysis. Englewood Cliffs, New Jersey: Prentice Hall.

Lehmann, E.L. (1966), "Some Concepts of Dependence," Annals of Statistics, 37, 1137-1153.

Linfoot, E.H. (1957), "An Informational Measure of Correlation," Information and Control, 1, 85-89.

Loeve, M. (1977), Probability Theory I, New York: Springer-Verlag.

_____ (1978), Probability Theory II, New York: Springer-Verlag.

Loftsgaarden, D.O., and Quesenberry, C.P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," Annals of Mathematical Statistics, 36, 1049-1051.

Mack, Y.P., and Rosenblatt, M. (1979), "Multivariate k-Nearest Neighbor Density Estimates," Journal of Multivariate Analysis, 9, 1-15.

Mardia, K.V. (1970), Families of Bivariate Distributions, London: Charles Griffin and Company.

Marshall, A.W., and Olkin, I. (1967), "A Multivariate Exponential Distribution," Journal of the American Statistical Association, 62, 30-44.

Moore, David S., and Yackel, James (1977), "Large Sample Properties of Nearest Neighbor Density Function Estimators," Statistical Decision Theory and Related Topics, New York: Academic Press.

Morrison, Donald F. (1976), Multivariate Statistical Methods, New York: McGraw-Hill.

Nadaraya, E.A. (1964), "On Estimating Regression," Theory of Probability and its Applications, 9, 141-142.

Newton, H. Joseph (1979), "The TIMESBOARD Time Series Analysis Computing Library," Institute of Statistics, Texas A and M University, College Station, Texas.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. (1975) SPSS: Statistical Package for the Social Sciences, 2nd ed., New York: McGraw-Hill.

Parzen, Emanuel (1959), "Statistical Inference on Time Series by Hilbert Space Methods I," Technical Report No. 23, Statistics Department, Stanford University.

_____ (1961), "Regression Analysis of Continuous Parameter Time Series," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, Berkeley: University of California Press.

_____ (1962), "On Estimation of a Probability Density Function and Mode," Annals of Mathematical Statistics, 33, 1065-1076.

_____ (1967), Time Series Analysis Papers, San Francisco: Holden-Day.

_____ (1977), "Nonparametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for White

Noise," Technical Report No. 47, Statistical Science Division, State University of New York at Buffalo.

_____ (1979), "Nonparametric Statistical Data Modeling," Journal of the American Statistical Association, 74, 105-131.

_____ (1982), "Quantiles, Parametric-Select Density Estimation, and Bi-Information Parameter Estimators," mimeograph, Institute of Statistics, Texas A and M University, College Station, Texas.

_____, and Anderson, Scott (1980), "ONESAM, A Computer Program for Nonparametric Data Analysis and Density Estimation," Technical Report B-1, Institute of Statistics, Texas A and M University, College Station, Texas.

Powell, M.J.D. (1981), Approximation Theory and Methods, Cambridge: Cambridge University Press.

Prihoda, Thomas J. (1981), "A Generalized Approach to the Two Sample Problem: The Quantile Approach," unpublished Ph.D. dissertation, Institute of Statistics, Texas A and M University.

Puri, M.L., Sen, P.K., and Gokhale, D.V. (1970), "On a Class of Rank Order Tests for Independence in Multivariate Distributions," Sankhya: Series A, 32, 271-298.

Rainville, Earl D. (1960), Special Functions, New York: The MacMillan Company.

Randles, Ronald H., and Wolfe, Douglas A. (1979), Introduction to the Theory of Nonparametric Statistics, New York: John Wiley and Sons.

Rao, C.R. (1973), Linear Statistical Inference and its Applications, New York: John Wiley and Sons.

Rosenblatt, M. (1956), "Remarks on Some Non-parametric Estimates of a Density Function," Annals of Mathematical Statistics, 27, 832-837.

Royden, H.L. (1968), Real Analysis, New york: MacMillan.

Ruymgaart, F.H. (1974), "Asymptotic Normality of Non-parametric Tests for Independence, Annals of Statistics, 2, 892-910.

Ryan, T.A., Joiner, B.L., and Ryan, B.F. (1975), Student Handbook for the MINITAB Statistical Computing System, Statistics Department, Pennsylvania State University.

SAS/GRAPH User's Guide (1981), SAS Institute Inc., Cary, North Carolina.

SAS User's Guide (1979), SAS Institute Inc., Cary, North Carolina.

Schwartz, S.C. (1967), "Estimation of a Probability Density by an Orthogonal Series," Annals of Mathematical Statistics, 38, 1261-1265.

Scott, D.W., and Factor, L.E. (1981), "Monte Carlo Study of Three Data-based Nonparametric Probability Density Estimators," Journal of the American Statistical Association, 76, 9-15.

Scott, D.W., Gotto, A.M., Cole, J.S., and Gorry, G.A. (1978), "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease - A Study of 371 Males with Chest Pain," Journal of Chronic Diseases, 31, 337-345.

Serfling, Robert J. (1980), Approximation Theorems of Mathematical Statistics, New York: John Wiley and Sons.

Shannon, C.E. (1948), "A Mathematical Theory of Communication," Bell System Technical Journal, 27, 379-423, 623-656.

Sillitto, G.P. (1969), "Derivation of Approximants to the Inverse Distribution Function of a Continuous Univariate Population from the Order Statistics of a Sample," Biometrika, 56, 641-650.

Silverman, Bernard (1978), "Weak and Strong Consistency of the Kernel Estimate of a Density and its Derivatives," Annals of Statistics, 6, 177-184.

_____ (1980), "Density Estimation: Are Theoretical Results Useful in Practice?" Asymptotic Theory of Statistical Tests and Estimation, I.M. Chakravarti, ed., New York: Academic Press.

Spearman, C. (1904), "The Proof and Measurement of Association between Two Things," American Journal of Psychology, 15, 72-101.

Stute, Winfried (1982), "The Oscillation Behaviour of Empirical Processes," Annals of Probability, 10, 86-107.

Taniguchi, Masanobu (1980), "On Estimation of the Integrals of Certain Functions of Spectral Density," Journal of Applied Probability, 17, 73-83.

Tapia, Richard A., and Thompson, James R. (1978), Nonparametric Probability Density Estimation, Baltimore: The Johns Hopkins University Press.

Tartar, M.E., and Kronmal, R.A. (1970), "On Multivariate Density Estimates Based on Orthogonal Expansions," Annals of Mathematical Statistics, 41, 718-722.

_____ (1976), "An Introduction to the Implementation and Theory

of Nonparametric Density Estimation," American Statistician, 30, 105-112.

Tartar, M.E., and Silvers, A. (1975), "Implementation and Applications of Bivariate Gaussian Mixture Decomposition," Journal of the American Statistical Association, 70, 47-55.

Tukey, John W. (1977), Exploratory Data Analysis, Reading, Massachusetts: Addison-Wesley Publishing Company.

Vasicek, Oldrich (1976), "A Test for Normality Based on Sample Entropy," Journal of the Royal Statistical Society, Series B, 38, 54-59.

Wahba, Grace (1971), "A Polynomial Algorithm for Density Estimation," Annals of Mathematical Statistics, 42, 1870-1886.

Walter, G.G. (1977), "Properties of Hermite Series Estimates of Probability Densities," Annals of Statistics, 5, 1258-1264.

Watson, G.S. (1964), "Smooth Regression Analysis," Sankhya: Series A, 26, 359-372.

_____, and Leadbetter, M.R. (1963), "On Estimating a Probability Density," Annals of Mathematical Statistics, 34, 480-491.

Wertz, Wolfgang (1978), Statistical Density Estimation: A Survey, Gottingen und Zurich: Vandenhoek and Ruprecht.

APPENDIX A

A.1 *Subprogram MIDEN*

```
      SUBROUTINE MIDEN(X,N,A,B,CAPT,IMOD)
C*****************************************************************
C
C   SUBROUTINE TO PERFORM NONPARAMETRIC DENSITY ESTIMATION USING
C   LEAST SQUARES REGRESSION MINIMUM INFORMATION TECHNIQUE.
C   LEGENDRE POLYNOMIALS ARE EMPLOYED FOR THE ORTHOGONAL EXPANSION
C
C   INPUT: X,N - DATA AND SAMPLE SIZE
C          A,B - MIN AND MAX DATA VALUES
C          CAPT - LABEL FOR X
C
C   OUTPUT: PLOTS AND DESCRIPTIVE STATISTICS
C
C   SUBPROGRAMS CALLED: PLOTXY,FTERP,MAX,MIN,
C                       SEQREG,LEGP,NNDEN,SWEEPD,RELMIN,CLPLT1
C
C*****************************************************************
      DIMENSION CAPT(20),X(N),Y(150),FNN(150),FMI(150)
      DIMENSION CBAR(13),AIC(13)
      DIMENSION THETA(12),P(3,12),COF(12,12),IORD(12),BEST(78,3),
     +RVAR(12),NVAR(12),COFF(11),MORD(5),XNORD(5)
      DOUBLE PRECISION COV(13,13)
      DATA NAIC/4H AIC/
      DATA NTHT/4HTHET/
      DATA NCOF/4H COF/
      DATA NAMX/4H  X /
      DATA NAMFN/4H FNN/
      DATA NAMFM/4H FMI/
      DATA NRVR/4HRVAR/
      DATA IORD/2,3,4,5,6,7,8,9,10,11,12,1/
      DATA MORD/2,3,6,11,8/
      DATA XNORD/4H=2  ,4H=3  ,4H=6  ,4H=11 ,4H=AIC/
      DATA XMN/4HM(N)/
      CAPT(13)=XMN
      XN=FLOAT(N)
      MN=10
      RANGE=B-A
      IDEG=0
      CALL LEGP(12,P,IDEG)
      DO 40 K=1,3
         DO 40 J=1,12
         COF(K,J)=P(K,J)
 40   CONTINUE
      DO 50 K=4,12
```

```
      CALL LEGP(12,P,IDEG)
         DO 50 J=1,12
         COF(K,J)=P(3,J)
50    CONTINUE
      DO 240 ITER1=1,2
      DO 25 I=1,N
      Y(I)=X(I)
25    CONTINUE
      IF(ITER1.EQ.2) MN=20
      CALL NNDEN(X,Y,N,N,MN,RANGE,FNN)
      CALL PLOTXY(Y,FNN,100,CAPT,NAMX,NAMFN,1)
      DO 55 I=1,N
      FNN(I)=ALOG(FNN(I))
55    CONTINUE
      DO 90 K=1,12
      CBAR(K)=0.0
         DO 80 J=1,K
         COV(K,J)=0.0
         DO 70 I=1,N
         ARGK=COF(K,1)
         ARGJ=COF(J,1)
         IF(K.EQ.1) GO TO 61
            DO 60 L=2,K
      ARGK=ARGK+COF(K,L)*((2.0*(X(I)-A)/RANGE-1.0)**(L-1))
      ARGJ=ARGJ+COF(J,L)*((2.0*(X(I)-A)/RANGE-1.0)**(L-1))
60          CONTINUE
61    CONTINUE
         IF(J.EQ.1)  CBAR(K)=CBAR(K)+ARGK
         COV(K,J)=COV(K,J)+ARGK*ARGJ
70       CONTINUE
      IF(J.EQ.1)  CBAR(K)=CBAR(K)/XN
      COV(K,J)=COV(K,J)/XN-CBAR(K)*CBAR(J)
      IF(K.NE.J)  COV(J,K)=COV(K,J)
80    CONTINUE
90    CONTINUE
      CBAR(13)=0.0
      DO 120 J=1,12
      COV(13,J)=0.0
         DO 110 I=1,N
         ARGJ=COF(J,1)
            DO 100 L=2,J
      ARGJ=ARGJ+COF(J,L)*((2.0*(X(I)-A)/RANGE-1.0)**(L-1))
100         CONTINUE
         IF(J.EQ.1)  CBAR(13)=CBAR(13)+FNN(I)
         COV(13,J)=COV(13,J)+ARGJ*FNN(I)
110      CONTINUE
      IF(J.EQ.1)  CBAR(13)=CBAR(13)/XN
      COV(13,J)=COV(13,J)/XN-CBAR(J)*CBAR(13)
      COV(J,13)=COV(13,J)
120   CONTINUE
      COV(13,13)=0.0
      DO 130 I=1,N
```

```
      COV(13,13)=COV(13,13)+FNN(I)*FNN(I)
130   CONTINUE
      COV(13,13)=COV(13,13)/XN-CBAR(13)*CBAR(13)
      DO 140 I=1,N
      FNN(I)=EXP(FNN(I))
140   CONTINUE
      CALL SEQREG(COV,13,12,IORD,BEST,78,RVAR,NVAR,NIVN)
      CALL CLPLT1(RVAR,NIVN,1,NRVR,41,1)
      NV1=NIVN+1
      AIC(1)=-1.0/XN
      AIC(2)=ALOG(RVAR(1))+2.0/XN
      DO 150 I=2,NIVN
      AIC(I+1)=ALOG(RVAR(I))+2.0*FLOAT(I)/XN
150   CONTINUE
      DEC=1./XN
      CALL RELMIN(AIC,NV1,DEC,MIN1,MIN2,NAIC)
      MIN1=MIN1-1
      MIN2=MIN2-1
      MORD(5)=MIN1
      DO 230 ITER=1,5
      KORD=MORD(ITER)
      CAPT(14)=XNORD(ITER)
      K=NVAR(KORD)
      DO 160 I=1,KORD
      THETA(I)=BEST(K,2)
      K=K+1
160   CONTINUE
      CALL CLPLT1(THETA KORD,1,NTHT,41,1)
      DO 180 K=2,KORD
      COFF(K-1)=0.0
      DO 170 J=2,KORD
      COFF(K-1)=COFF(K-1)+COF(J,K)*THETA(J-1)
170   CONTINUE
180   CONTINUE
      CALL CLPLT1(COFF,KORD-1,1,NCOF,41,1)
      CALL EQSPY(A,B,100,Y)
      FSUM=0.0
      FMODE=0.
      XMODE=999.
      FMEAN=0.0
      FVAR=0.0
      DO 200 I=1,100
      FMI(I)=0.0
      DO 190 L=2,KORD
      FMI(I)=FMI(I)+COFF(L-1)*((2.0*(Y(I)-A)/RANGE-1.0)**(L-1))
190   CONTINUE
      FMI(I)=EXP(FMI(I))
      FSUM=FSUM+FMI(I)
      FMEAN=FMEAN+Y(I)*FMI(I)
      FVAR=FVAR+Y(I)*Y(I)*FMI(I)
      IF(FMODE.GT.FMI(I)) GO TO 200
      FMODE=FMI(I)
```

```
      XMODE=Y(I)
  200 CONTINUE
      FSUM=FSUM*RANGE/100.0
      FMEAN=FMEAN*RANGE/(FSUM*100.0)
      FMODE=FMODE/FSUM
      FVAR=FVAR*RANGE/(FSUM*100.0)
      FVAR=FVAR-FMEAN*FMEAN
      DTEST=0.0
      DO 210 I=1,N
      FMI(I)=FMI(I)/FSUM
      DTEST=DTEST+ABS(FMI(I)-FNN(I))
  210 CONTINUE
      DTEST=DTEST/XN
      WRITE(6,220) DTEST
  220 FORMAT(/,10X,'AVE OF SUM OF ABS(FMI-FNN)=',F10.4,/)
      WRITE(6,284)
  284 FORMAT(1H1,7(/))
      WRITE(6,290) FMEAN,FVAR,FMODE,XMODE,FSUM
  290 FORMAT(///,13X,'+',108(1H-),'+',/,13X,'|   PARAMETER ',
     +'FUNCTIONALS: MEAN =',F10.4,',   VARIANCE =',F10.4,',   MODE =',
     +F10.4,' AT X = ',F10.4,'    |',/,13X,'|   INTEGRATING FACTOR =',
     +F10.4,75X,'|',/,13X,'+',108(1H-),'+',/)
      CALL PLOTXY(Y,FMI,100,CAPT,NAMX,NAMFM,1)
  230 CONTINUE
  240 CONTINUE
      RETURN
      END
```

A.2 Subprogram CMPDEN

```
      SUBROUTINE CMPDEN(W,N,CAPT,THETA,INDW,NVW,ISORT)
C*****************************************************************
C
C   THIS SUBPROGRAM COMPUTES A SMOOTHED DENSITY QUANTILE
C   FUNCTION BASED ON ORTHOGONAL EXPANSION IN TERMS OF
C   COMPLEX EXPONENTIALS. A NEAREST NEIGHBOR ESTIMATE IS
C   OBTAINED AS A RAW DENSITY QUANTILE AND THEN A SEQUENTIAL
C   REGRESSION ROUTINE IS APPLIED WITH THE RAW D-Q TREATED
C   AS A DEPENDENT VARIABLE. A COMPLEX SWEEP OPERATOR IS THEN
C   USED TO OBTAIN COEFFICIENTS AND RESIDUAL VARIANCES FOR
C   VARIOUS ORDERS, AND FINALLY PARZEN'S CAT CRITERION IS
C   USED TO CHOOSE THE 'BEST' EXPANSION.
C
C   INPUT: W - RAW DATA
C          N - SAMPLE SIZE
C          ISORT - 0 IF W SORTED, 1 OTHERWISE.
C
C   OUTPUT: NVW - ORDER OF EXPANSION CHOSEN (NUMBER OF INDEP. VAR.)
```

```
C           THETA - COEFFICIENTS IN EXPANSION
C           INDW - VECTOR OF INDICES CORRESPONDING TO COEFFICIENTS
C                  CHOSEN
C
C   SUBROUTINES CALLED: CSQREG,CSWEEP
C
C*****************************************************************
      DIMENSION W(N),FMC(200),THETA(21),INDW(21),BEST(231,3),
     +RVAR(21),NVAR(21),IORDI(21),IORD(21),FNN(200),T(200),FHO(200)
      DIMENSION CAPT(20),MORD(3),CORD(3)
      COMPLEX PHI(22,22),PHIBAR(21),ZARG,CEXP,CONJG,CMPLX
      DATA IORDI/11,10,12,9,13,8,14,7,15,6,16,5,17,4,18,3,19,2,20,1,21/
      DATA LVAR,LTHT/4HRVAR,4HTHET/
      DATA NAMX/4H    X/
      DATA NAMFN,NAMFC/4H FNN,4H FMC/
      DATA NAMM/4H FHO/
      DATA MORD/2,8,15/
      DATA CORD/4H=2  ,4H=8  ,4H=15 /
      DATA CMN/4HM(N)/
      CAPT(13)=CMN
      DO 1 I=1,21
      IORD(I)=IORDI(I)
    1 CONTINUE
      DO 5 I=1,N
      T(I)=W(I)
    5 CONTINUE
      N2=N-2
      M=21
      M1=M+1
      MID=11
      TWOPI=8.0*ATAN(1.0)
      DENOM=1./FLOAT(N+1)
      IF(ISORT.EQ.0) GO TO 10
      CALL GRD(T,N)
      CALL ORD(W,N)
   10 CONTINUE
      A=T(1)-DENOM
      B=T(N)+DENOM
      RANGE=B-A
C
C   COMPUTE NEAREST NEIGHBOR DENSITY ESTIMATE
C
      DO 200 ITER1=1,2
      MN=8
      IF(ITER1.EQ.2) MN=15
      CAPT(14)=CORD(2)
      IF(ITER1.EQ.2) CAPT(14)=CORD(3)
      CALL NNDEN(T,W,N,N,MN,RANGE,FNN)
      CALL PLOTXY(T,FNN,N,CAPT,NAMX,NAMFN,1)
      FBAR=0.0
      DO 40 I=3,N2
      FBAR=FBAR+FNN(I)
```

```
 40   CONTINUE
      FBAR=FBAR/FLOAT(N-4)
      DO 60 I=1,M
      II=I-MID
      PHIBAR(I)=CMPLX(0.,0.)
      DO 50 K=3,N2
      ARG=TWOPI*FLOAT(II)*(T(K)-A)/RANGE
      ZARG=CMPLX(0.,ARG)
      PHIBAR(I)=PHIBAR(I)+CEXP(ZARG)
 50   CONTINUE
      PHIBAR(I)=PHIBAR(I)/FLOAT(N-4)
 60   CONTINUE
      DO 90 I=1,M
      DO 80 J=1,I
      II=I-MID
      JJ=J-MID
      PHI(I,J)=CMPLX(0.,0.)
      DO 70 K=3,N2
      ARG=TWOPI*(FLOAT(II)-FLOAT(JJ))*(T(K)-A)/RANGE
      ZARG=CMPLX(0.,ARG)
      PHI(I,J)=PHI(I,J)+CEXP(ZARG)
 70   CONTINUE
      PHI(I,J)=PHI(I,J)/FLOAT(N-4)-PHIBAR(I)*CONJG(PHIBAR(J))
      PHI(J,I)=CONJG(PHI(I,J))
 80   CONTINUE
 90   CONTINUE
      DO 110 I=1,M
      II=I-MID
      PHI(M1,I)=CMPLX(0.,0.)
      DO 100 K=3,N2
      ARG=-TWOPI*FLOAT(II)*(T(K)-A)/RANGE
      ZARG=CMPLX(0.,ARG)
      PHI(M1,I)=PHI(M1,I)+FNN(K)*CEXP(ZARG)
100   CONTINUE
      PHI(M1,I)=PHI(M1,I)/FLOAT(N-4)-FBAR*CONJG(PHIBAR(I))
      PHI(I,M1)=CONJG(PHI(M1,I))
110   CONTINUE
      PHI(M1,M1)=CMPLX(0.,0.)
      DO 120 K=3,N2
      PHI(M1,M1)=PHI(M1,M1)+FNN(K)*FNN(K)
120   CONTINUE
      PHI(M1,M1)=PHI(M1,M1)/FLOAT(N-4)-FBAR*FBAR
      CALL CSQREG(PHI,22,M,IORD,BEST,231,RVAR,NVAR,NIVN)
      CALL CLPLT1(RVAR,NIVN,1,LVAR,41,1)
      NV1=NIVN+1
      DO 190 ITER=1,3
      NVW=MORD(ITER)
      CAPT(14)=CORD(ITER)
      K1=NVAR(NVW)
      DO 140 K=1.NVW
      INDW(K)=IFIX(BEST(K1,1)+0.5)
      THETA(K)=BEST(K1,2)
```

```
      K1=K1+1
  140 CONTINUE
      WRITE(6,150)
  150 FORMAT(1H1)
      CALL CLPLT1(THETA,NVW,1,LTHT,41,1)
      FSUM=0.0
      DO 170 I=1,50
      FMC(I)=0.0
      T(I)=A+FLOAT(I)*RANGE/50.0
      DO 160 K=1,NVW
      ARG=TWOPI*FLOAT(INDW(K)-11)*(T(I)-A)/RANGE
      FMC(I)=FMC(I)+THETA(K)*COS(ARG)
  160 CONTINUE
      FMC(I)=EXP(FMC(I))
      FSUM=FSUM+FMC(I)
  170 CONTINUE
      FSUM=FSUM*RANGE/50.0
      DO 180 I=1,50
      FMC(I)=FMC(I)/FSUM
      FHO(I)=.5*XNORM(0.,1.,T(I))+.5*XNORM(3.,.5,T(I))
  180 CONTINUE
      CALL SEDIAG(FHO,FMC,50,RANGE,XISE,XMAXD,XMSE)
      CALL PLTXYZ(T,FMC,FHO,50,CAPT,NAMX,NAMFC,NAMM,XISE,XMAXD,XMSE)
  190 CONTINUE
  200 CONTINUE
      RETURN
      END
```

A.3 Two Step FORTRAN-SAS Program Duplicating MIDEN

In the following listing, items appearing in lowercase represent options depending on the system and the intended application.

```
// job card
// optional operating system cards
//STEP1 EXEC FORTX,REGION=512K        <--- one step FORTRAN procedure
//FT01F001 DD DSN=WYL.scratch file name
//SYSLIB DD
// DD
// DD
// DD
// DD name of user subroutine library (TIMESBOARD in this case)
//SOURCE DD *
C
C   PROGRAM TO PERFORM NONPARAMETRIC DENSITY ESTIMATION USING
C   LEAST SQUARES REGRESSION. DATA SET WRITTEN FOR USE BY SAS GLM.
```

```
C   SEE WOODFIELD DISSERTATION FOR MORE INFORMATION.
C
      DIMENSION CAPT(20),L(5),X(150),FNN(150),Y(150)
      DIMENSION THETA(12),P(3,12),COF(12,12)
C
C   READ DATA SET
C
      READ(5,10) CAPT
   10 FORMAT(20A4)
      READ(5,20)  N,L
   20 FORMAT(I5,4X,5A4)
      READ(5,L)  (X(I),I=1,N)
      DO 25 I=1,N
      Y(I)=X(I)
   25 CONTINUE
      A=X(1)
      B=A
      DO 30 I=2,N
      IF(X(I).GT.B)  B=X(I)
      IF(X(I).LT.A)  A=X(I)
   30 CONTINUE
      A=A-1.0/FLOAT(N)
      B=B+1.0/FLOAT(N)
      RANGE=B-A
      IDEG=0
      CALL LEGP(12,P,IDEG)
      DO 40 K=1,3
         DO 40 J=1,K
         COF(K,J)=P(K,J)
   40 CONTINUE
      DO 50 K=4,12
      CALL LEGP(12,P,IDEG)
         DO 50 J=1,K
         COF(K,J)=P(3,J)
   50 CONTINUE
      CALL NNDEN(X,Y,N,N,10,FNN)
      DO 70 I=1,N
         DO 60 K=2,12
         THETA(K)=COF(K,1)
             DO 60 J=2,K
      THETA(K)=THETA(K)+COF(K,J)*((2.0*(X(I)-A)/RANGE-1.0)**(J-1))
   60    CONTINUE
      WRITE(1,65)  FNN(I),(THETA(K),K=2,12)
   65 FORMAT(1X,7F10.5,/,1X,5F10.5)
   70 CONTINUE
      STOP
      END
      SUBROUTINE LEGP(N,P,IDEG)
      .
      .
      code for subprogram LEGP
      .
```

```
      .
      SUBROUTINE NNDEN(X,Y,N,M,MN,FNN)
      .
      .
      code for subprogram NNDEN
      .
      .
 //SYSIN DD *
      data goes here
 //*
 //STEP2 EXEC SAS
 //ONE DD DSN=WYL. description of tape 1 above where output
                  of STEP1 was written
DATA TWO; INFILE ONE:
INPUT FNN X1-X6 #2 X7-X11;
Y=LOG(FNN);
CARDS;
PROC GLM;
MODEL Y = variable listing for variables to be included in model/ P;
OUTPUT OUT=NEW PREDICTED=YP;
DATA THREE; SET NEW;
FHAT=EXP(YP);
PROC PLOT DATA=THREE; PLOT FHAT*X1='*';
```

APPENDIX B

BISAM: A Program for Bivariate Data Analysis

```
C
C   PROGRAM BISAM
C
C*************************************************************
C
C   DRIVER PROGRAM FOR BIVARIATE DATA ANALYSIS
C
C   THIS PROGRAM PRODUCES SCATTER PLOTS, DESCRIPTIVE STATISTICS,
C   AND CORRELATION STATISTICS FOR A SET OF BIVARIATE DATA.
C   BIVARIATE DENSITY ESTIMATION IS PERFORMED ON THE DEPENDENCE
C   DENSITY USING MINIMUM BI-INFORMATION TECHNIQUES.
C   A 40X40 GRID OF DENSITY AND DENSITY QUANTILE VALUES IS
C   WRITTEN TO TAPES 1 THROUGH 3 FOR THE ORDERS 8, 24, AND 48
C   TO BE USED FOR GRAPHICAL OUTPUT USING SAS/GRAPH.
C   SEE WOODFIELD DISSERTATION FOR MORE INFORMATION.
C
C   INPUT: NTAPE - TAPE WHERE DATA SET RESIDES
C          (X,Y) - BIVARIATE DATA (INDIVIDUALLY, X FIRST)
C          MORD - MAXIMUM AUTOREGRESSIVE ORDER TO BE USED FOR
C                 UNIVARIATE AR DENSITY ESTIMATION (<=6)
C          IDQX,IDQY - NULL DISTRIBUTIONS FOR AUTOREGRESSIVE SMOOTHING
C          IPLT1 - 0--> NO SCATTER PLOTS
C                  1--> SCATTER PLOT OF DATA
C                  2--> SCATTER PLOT OF RANK TRANSFORMED DATA
C                  3--> BOTH SCATTER PLOTS
C          IPLT2 - 0--> NO AUTOREGRESSIVE DENSITY PLOTS
C                  1--> BEST ORDER AR DENSITY PLOTS
C          IPLT3 - 0--> NO QUANTILE BOX PLOTS
C                  1--> QUANTILE BOX PLOTS FOR BOTH X AND Y
C          IDST - 0--> NO UNIVARIATE DESCRIPTIVE STATISTICS
C                 1--> UNIVARIATE DESCRIPTIVE STATISTICS FOR X AND Y
C          KDEL - MAXIMUM NUMBER OF EXTREME POINTS TO EXCLUDE FROM
C                 BIVARIATE ANALYSIS
C
C   SUBPROGRAMS CALLED: DATAIN,RANK,ORD2,PEARSN,SPRMN,PPLOT,TRIM,
C                       KENDAL,CMPINF,CPTENT,RELMIN,MIN,PLOTXY,FTERP,
C                       MINMAX,CSQREG,CSWEEP,AUTDEN,ORD,QHLIN,
C                       QTOFQ,WSPACE,FORIER,AUTORG,PARZ,AREST,
C                       FQFNC,MDNRIS,QFIND,MAX,CLPLT1,DESTAT,QPLOT,
C                       EXPAND
C
C
C*************************************************************
C
```

```
      COMMON X(500),Y(500),RANKX(500),RANKY(500)
      DIMENSION L(5),LABY(20),LABX(20),T(500,2),HD(4)
      DIMENSION CHAR(5),XNAME(20),YNAME(20),CAPT(20),CRNK(6)
      DIMENSION W(500),SA(1003),GE(1003)
      EQUIVALENCE (T(1,1),X(1))
      EQUIVALENCE (T(1,2),Y(1))
      DATA NOUT,NIN/6,5/
      DATA CHAR/1H*,1H+,1HX,1H#,1H./
      DATA XNAME/10*1H ,1HX,9*1H /
      DATA YNAME/10*1H ,1HY,9*1H /
      DATA CAPT/4HSCAT,4HTER ,4HPLOT,4H OF ,4HX VS,4H. Y ,14*4H    /
      DATA CRNK/4H- RA,4HNK T,4HRANS,4HFORM,4HATIO,4HN   /
C
      WRITE(NOUT,1)
    1 FORMAT(1H1)
      READ(NIN,10) NTAPE,IDQX,IDQY,MORD,IPLT1,IPLT2,IPLT3,IDST,KDEL
   10 FORMAT(9I5)
      WRITE(NOUT,20)
   20 FORMAT(//,10X,20(4H****),/,10X,'*',78X,'*',/,10X,'*   BISAM ',
     +'- BIVARIATE DATA ANALYSIS USING FOURIER EXPANSIONS',19X,'*',
     +/,10X,'*            AND QUANTILE TECHNIQUES',44X,'*',/,10X,'*',
     +78X,'*',/,10X,20(4H****))
      CALL DATAIN(NTAPE,X,NX,L,LABX)
      CALL DATAIN(NTAPE,Y,NY,L,LABY)
      N=NX
      IF(NX.EQ.NY) GO TO 40
      WRITE(NOUT,30) LABX,LABY
   30 FORMAT(1H ,10X,20A4/,10X,20A4,//,10X,'SAMPLE SIZES NOT EQUAL.',
     +' BIVARIATE ANALYSIS INAPROPRIATE. EXECUTION TERMINATED.')
      STOP
   40 WRITE(NOUT,50) LABX,LABY,N
   50 FORMAT(1H ,9X,20A4/,10X,20A4,//,10X,'N=',I5)
      IF((IPLT1.EQ.1).OR.(IPLT1.EQ.3))
     +CALL PPLOT(X,Y,500,N,1,CHAR,CAPT,XNAME,YNAME,0)
      WRITE(NOUT,1)
C
C   ORDER BIVARIATE DATA BY X VALUES
C
      CALL ORD2(T,N,500)
      IF(IDST.EQ.0) GO TO 58
      NN1=2*N+1
      DO 51 I=1,N
      W(I)=X(I)
   51 CONTINUE
      CALL DESTAT(W,N,LABX,IPLT3,XMED,SA,GE,NN1)
      DO 52 I=1,N
      W(I)=Y(I)
   52 CONTINUE
      CALL ORD(W,N)
      CALL DESTAT(W,N,LABY,IPLT3,YMED,SA,GE,NN1)
C
C   TRIM DATA SET OF AT MOST KDEL EXTREME POINTS
```

```
C
  58  CALL TRIM(X,Y,XMED,YMED,KDEL,N,NEWN)
      N=NEWN
C
C   OBTAIN RANKS OF X AND Y VALUES
C
      CALL RANK(X,N,RANKX)
      CALL RANK(Y,N,RANKY)
C
C   COMPUTE CORRELATION COEFFICIENTS
C
      CALL SPRMN(N,RHO,SUMD)
      CALL KENDAL(N,TAUA,TAUB,SOMER,NC,ND,NIND,NDEP,NPAIRS)
      CALL PEARSN(N,R)
      CALL CMPINF(N,MORD,IDQX,IDQY,IPLT2,HD)
C
C   WRITE VALUES OF CORRELATION COEFFICIENTS
C
      IF((NIND.EQ.0).AND.(NDEP.EQ.0)) GO TO 59
      WRITE(NOUT,55) NIND,NDEP
  55  FORMAT(//,10X,'TIES IN X =',I4,', TIES IN Y =',I4,//)
  59  WRITE(NOUT,50) LABX,LABY,N
      WRITE(NOUT,60)
  60  FORMAT(10X,'   PEARSON  SPEARMAN KENDALL A KENDALL B',
     +'    SOMER D  H(D-TIL)      H(D8)     H(D24)     H(D48)',
     +/,10X,10(9H---------))
      WRITE(NOUT,70) R,RHO,TAUA,TAUB,SOMER,HD(4),HD(1),HD(2),HD(3)
  70  FORMAT(10X,9F10.4)
      DO 80 I=1,N
      X(I)=RANKX(I)/FLOAT(N+1)
      Y(I)=RANKY(I)/FLOAT(N+1)
  80  CONTINUE
      DO 90 I=1,6
      CAPT(I+6)=CRNK(I)
  90  CONTINUE
      IF((IPLT1.EQ.2).OR.(IPLT1.EQ.3))
     +CALL PPLOT(X,Y,500,N,1,CHAR,CAPT,XNAME,YNAME,0)
C
      STOP
      END
      SUBROUTINE CMPINF(N,MORD,IDQX,IDQY,IPLT2,HD)
C*************************************************************
C
C   SUBROUTINE TO COMPUTE COVARIANCE MATRIX OF COMPLEX
C   EXPONENTIAL "SUFFICIENT STATISTICS" TO BE USED IN
C   SEQUENTIAL REGRESSION ROUTINE TO OBTAIN "BEST REGRESSION"
C   MODELS FOR ORDERS 1 THROUGH M*M. VARIOUS ORDER DETERMINING
C   CRITERION ARE COMPUTED AND DISPLAYED VIA SUBROUTINE CPTENT.
C   THE BIVARIATE DENSITY QUANTILE IS FORMED BY TAKING THE PRODUCT
C   OF THE ESTIMATED DEPENDENCE DENSITY AND THE UNIVARIATE
C   AUTOREGRESSIVE ESTIMATORS.
C
```

```
C    INPUT: RANKX,RANKY - VECTORS CONTAINING RANKS OF X AND CORANKS
C                         OF Y
C           X,Y - BIVARIATE DATA
C           N - SAMPLE SIZE
C           IDQX,IDQY - NULL UNIVARIATE DENSITY INDICATORS
C           MORD - MAXIMUM ORDER FOR AR DENSITY EST. PROCEDURE
C           IPLT2 - PLOTTING OPTION FOR UNIVARIATE DENSITY QUANTILES
C
C    OUTPUT: PHI - COVARIANCE MATRIX
C            FQX,FQY - UNIVARIATE DENSITY QUANTILE FUNCTIONS
C            DQHAT - BIVARIATE DENSITY QUANTILE FUNCTION
C            HD - VECTOR OF ENTROPY ESTIMATORS: 1 - ORDER 8
C                                               2 - ORDER 24
C                                               3 - ORDER 48
C                                               4 - RAW (FROM D-TILDA)
C
C    NOTE: FQX,FQY ARE NOT PASSED BACK TO THE CALLING PROGRAM.
C          ALSO, CRITERION FUNCTIONS ARE PLOTTED BUT NOT PASSED
C          BACK TO THE CALLING PROGRAM.
C
C    SUBPROGRAMS CALLED: CSQREG,CPTENT,PLOTXY,FTERP,AUTDEN,RELMIN,MIN
C
C*****************************************************************
C
      COMMON X(500),Y(500),RANKX(500),RANKY(500)
      DIMENSION IND(97),RADSQ(500),IORD(49),
     +HD(4),DTIL(500)
      DIMENSION MENT(3)
      COMPLEX ARGM(13,13),PHI(50,50),CEXP,CONJG,CMPLX,ZARG
      COMPLEX ALPHX(5),ALPHY(5),COF(97)
      DATA IORD/25,24,18,17,26,32,19,31,33,23,11,16,10,30,12,9,
     +27,39,20,38,34,40,13,37,41,22,4,15,3,29,5,8,2,36,6,1,28,
     +46,21,45,35,47,14,44,42,48,7,43,49/
      DATA MENT/8,24,48/
      REAL LGDHAT
      IF(N.GT.29) GO TO 20
      WRITE(6,10) N
   10 FORMAT(10X,'SAMPLE SIZE ',I2,' IS TOO SMALL. CMPINF SKIPPED.')
      RETURN
C
C   SET VALUES OF CONSTANTS
C
   20 N2=N-2
      DO 21 I=1,4
   21 HD(I)=999.0
C
C   FOR THIS VERSION USING COMPLEX SEQUENTIAL REGRESSION THE
C   MAXIMUM APPROXIMATING ORDER IS SET AT 7.
C
      M=7
      L=MOD(M,2)
      ML=(M-L)/2
```

```
      IF(L.EQ.0)  M=M+1
      M2=2*M-1
      MM=M*M
      M1=MM+1
      MM1=MM-1
      DENOM=1.0/FLOAT(N+1)
      TWOPI=8.0*ATAN(1.0)
      PI=TWOPI/2.0
C
C   COMPUTE NEAREST NEIGHBOR DENSITY ESTIMATE AND RAW ESTIMATE
C   OF ENTROPY
C
      HDO=0.0
      DO 30 I=3,N2
      DO 25 J=1,N
      RADSQ(J)=(RANKX(I)-RANKX(J))**2+(RANKY(I)-RANKY(J))**2
   25 CONTINUE
      DO 26 K=1,5
      CALL MIN(RADSQ,N,RMIN,INDR)
      RADSQ(INDR)=FLOAT(2*N*N)
   26 CONTINUE
      VKJ=RMIN*DENOM*DENOM*PI
      IF(VKJ.EQ.0.0)  VKJ=0.5*DENOM*DENOM*PI
C
C   DTIL IS ALOG(DTIL)
C
      DTIL(I)=ALOG(5.0/(FLOAT(N+1)*VKJ))
      HDO=HDO-DTIL(I)
   30 CONTINUE
      HD(4)=HDO/FLOAT(N-4)
C
C   COMPUTE MATRIX OF EXPONENTIAL CROSS-PRODUCTS TO BE USED FOR
C   COVARIANCE COMPUTATIONS
C
      DO 50 I=1,M2
      I1=I-M
      DO 50 J=1,M2
      J1=J-M
      ARGM(I,J)=CMPLX(0.0,0.0)
      DO 40 K=3,N2
      ARG=TWOPI*(FLOAT(I1)*RANKX(K)+FLOAT(J1)*RANKY(K))*DENOM
      ZARG=CMPLX(0.,ARG)
      ARGM(I,J)=ARGM(I,J)+CEXP(ZARG)
   40 CONTINUE
      ARGM(I,J)=ARGM(I,J)/FLOAT(N-4)
   50 CONTINUE
C
C   COMPUTE COVARIANCE MATRIX
C
      DO 60 IN=1,MM
      I=IN-1
      I2=MOD(I,M)
```

```
      I1=(I-I2)/M+M-ML
      I2=I2+M-ML
      DO 55 JN=1,IN
      J=JN-1
      J2=MOD(J,M)
      J1=(J-J2)/M
      I1=I1-J1+ML
      JJ=I2-J2+ML
      J1=J1+M-ML
      J2=J2+M-ML
      PHI(IN,JN)=ARGM(I1,JJ)-ARGM(I1,I2)*CONJG(ARGM(J1,J2))
      PHI(JN,IN)=CONJG(PHI(IN,JN))
   55 CONTINUE
   60 CONTINUE
C
C    COMPUTE LAST ROW OF COVARIANCE MATRIX
C
      DBAR=0.0
      DO 70 I=3,N2
      DBAR=DBAR+DTIL(I)
   70 CONTINUE
      DBAR=DBAR/FLOAT(N-4)
      DO 90 IN=1,MM
      I=IN-1
      I2=MOD(I,M)
      I1=(I-I2)/M-ML
      I2=I2-ML
      PHI(M1,IN)=CMPLX(0.0,0.0)
      DO 80 K=3,N2
      ARG=TWOPI*(FLOAT(I1)*RANKX(K)+FLOAT(I2)*RANKY(K))*DENOM
      ZARG=CMPLX(0.0,ARG)
      PHI(M1,IN)=PHI(M1,IN)+DTIL(K)*CONJG(CEXP(ZARG))
   80 CONTINUE
      PHI(M1,IN)=PHI(M1,IN)/FLOAT(N-4)-DBAR*CONJG(ARGM(I1+M,I2+M))
      PHI(IN,M1)=CONJG(PHI(M1,IN))
   90 CONTINUE
      PHI(M1,M1)=0.0
      DO 100 K=3,N2
      PHI(M1,M1)=PHI(M1,M1)+DTIL(K)*DTIL(K)
  100 CONTINUE
      PHI(M1,M1)=PHI(M1,M1)/FLOAT(N-4)-DBAR*DBAR
C
C    CALL ROUTINE CPTENT TO COMPUTE AND PLOT CRITERION FUNCTIONS AND
C    DETERMINE BEST AND SECOND BEST MODELS FOR D(U1,U2)
C
      CALL CPTENT(RANKX,RANKY,N,M,PHI,IORD,IND,COF,MENT,HD)
C
C    COMPUTE UNIVARIATE DENSITY ESTIMATES USING AUTOREGRESSIVE
C     TECHNIQUE
C
      WRITE(6,134)
  134 FORMAT(1H1)
```

```
      CALL AUTDEN(X,N,IDQX,IPLT2,MORD,ALPHX,RVARX,SIGX,NVX,0,4H  X )
      WRITE(6,134)
      CALL AUTDEN(Y,N,IDQY,IPLT2,MORD,ALPHY,RVARY,SIGY,NVY,1,4H  Y )
      WRITE(6,134)
      WRITE(6,135) NVX,NVY
  135 FORMAT(/,10X,'UNIVARIATE BEST ORDERS: NVX =',I3,', NVY =',I3)
      DO 260 ITER=1,3
      PSI=0.0
      ENT=0.0
      WRITE(6,136)
  136 FORMAT(/,19X,'U1',17X,'U2',15X,'DQHT',15X,'DHAT',/,2X,
     +19(4H----))
      DO 220 I=1,40
      DO 220 J=1,40
      U1=FLOAT(I)/41.
      U2=FLOAT(J)/41.
C
C   COMPUTE VALUES OF UNIVARIATE DENSITY-QUANTILE FUNCTIONS
C
      FQX=1.0
      IF(NVX.GT.0) FQX=AREST(U1,NVX,RVARX,ALPHX)
      FQX=FQFNC(U1,IDQX)/(FQX*SIGX)
      FQY=1.0
      IF(NVY.GT.0) FQY=AREST(U2,NVY,RVARY,ALPHY)
      FQY=FQFNC(U2,IDQY)/(FQY*SIGY)
C
C   COMPUTE BIVARIATE DENSITY QUANTILE BY FORMING PRODUCT
C   OF DEPENDENCE DENSITY AND AUTOREGRESSIVE ESTIMATORS
C
      LGDHAT=0.0
      KP=MENT(ITER)
      LOC=1
      IF(ITER.EQ.2) LOC=MENT(1)+1
      IF(ITER.EQ.3) LOC=MENT(1)+MENT(2)+1
      DO 200 K=1,KP
      I1=IND(LOC)-1
      I2=MOD(I1,M)
      I1=(I1-I2)/M-ML
      I2=I2-ML
      ARG=TWOPI*(FLOAT(I1)*U1+FLOAT(I2)*U2)
      ZARG=CMPLX(0.0,ARG)
      LGDHAT=LGDHAT+REAL(COF(LOC)*CEXP(ZARG))
      LOC=LOC+1
  200 CONTINUE
      IF(LGDHAT.GT.170.) RETURN
      IF(LGDHAT.LT.-20.) LGDHAT=-20.
      DHAT=EXP(LGDHAT)
      ENT=ENT-LGDHAT*DHAT
      PSI=PSI+DHAT
      DQHT=DHAT*FQX*FQY
      WRITE(ITER,210) U1,U2,DQHT,DHAT
  210 FORMAT(2X,4F19.10)
```

```
      IMOD=MOD(I,8)
      JMOD=MOD(J,8)
      IF((IMOD.EQ.0).AND.(JMOD.EQ.0)) WRITE(6,210) U1,U2,DQHT,DHAT
C
 220  CONTINUE
      PSI=PSI/1681.0
      ENT=ENT/1681.0
      HD(ITER)=ENT/PSI+ALOG(PSI)
      WRITE(6,224)
 224  FORMAT(//,10X,20(4H====),//)
      WRITE(6,225) MENT(ITER),PSI
 225  FORMAT(/,10X,'INTEGRATING FACTOR FOR ORDER ',I3,' IS ',F10.4)
      WRITE(6,230)
 230  FORMAT(//,10X,'COEFFICIENTS FOR BIVARIATE DEPENDENCE DENSITY',
     +//,12X,'NU1',2X,'NU2',2X,' REAL(COF)  IMAG(COF) ',/,10X,32(1H-))
      LOC=1
      IF(ITER.EQ.2) LOC=MENT(1)+1
      IF(ITER.EQ.3) LOC=MENT(1)+MENT(2)+1
      DO 250 I=1,KP
      I1=IND(LOC)-1
      I2=MOD(I1,M)
      I1=(I1-I2)/M-ML
      I2=I2-ML
      WRITE(6,240) I1,I2,COF(LOC)
 240  FORMAT(10X,2I5,2F10.4)
      LOC=LOC+1
 250  CONTINUE
      WRITE(6,224)
 260  CONTINUE
      RETURN
      END
      SUBROUTINE CPTENT(RANKX,RANKY,N,M,PHI,IORD,IND,COF,MENT,HD)
C*****************************************************************
C
C   SUBPROGRAM TO COMPUTE AND PLOT ENTROPY OF D-HAT.
C   THIS SUBROUTINE WILL ALSO COMPUTE THE CRITERION FUNCTION
C   AIC AND PRINT THE SMALLEST TWO RELATIVE MINIMA.
C   COEFFICIENTS FOR THE THREE ORDERS SPECIFIED IN MENT
C   WILL BE RETURNED IN COF WITH THE CORRESPONDING INDICES
C   IN IND.
C
C   INPUT: N,M - SAMPLE SIZE, UNIVARIATE MAXIMUM ORDER (M**2
C                USED FOR BIVARIATE MAX ORDER)
C          RANKX,RANKY - VECTORS OF RANKS AND CO-RANKS
C          PHI - COVARIANCE MATRIX
C          IORD - VECTOR OF ORDERED INDICES FOR SEQUENTIAL REGRESSION
C
C   AUXILLIARY: NVAR,RVAR,BEST - VECTORS AND MATRIX
C               FROM ROUTINE CSQREG
C
C   OUTPUT: COF,IND - SEE ABOVE
C
```

```
C   SUBROUTINES CALLED: PLOTXY,FTERP,MINMAX,RELMIN,MIN,CSQREG,CSWEEP
C
C*****************************************************************
C
      DIMENSION RANKX(N),RANKY(N),AIC(50),
     +MENT(3),IND(97),HD(4)
      DIMENSION IORD(49),NVAR(49),INDV(1225),RVAR(49)
      COMPLEX PHI(50,50),COF(97),BEST(1225)
      REAL LGDHAT
      MM=M*M
      MM1=MM-1
      L=MOD(M,2)
      ML=(M-L)/2
      TWOPI=8.0*ATAN(1.0)
C
C   CALL ROUTINE CSQREG TO PERFORM SEQUENTIAL REGRESSION ON PHI
C
      CALL CSQREG(PHI,50,MM,IORD,BEST,INDV,1225,RVAR,NVAR,NIVN)
      CALL CLPLT1(RVAR,NIVN,1,4HRVAR,41,1)
      NV1=NIVN+1
C
C   COMPUTE AIC CRITERION FUNCTION
C
      AIC(1)=-1./FLOAT(N)
      AIC(2)=ALOG(RVAR(1))+2.0/FLOAT(N)
      DO 30 I=2,NIVN
      AIC(I+1)=ALOG(RVAR(I))+2.*FLOAT(I)/FLOAT(N)
   30 CONTINUE
      DEC=1.0/FLOAT(N)
      WRITE(6,40)
   40 FORMAT(/,10X,'OUTPUT FROM RELMIN FOR ORDER DETERMINING ',
     +'CRITERION. (SUBTRACT ONE FOR TRUE ORDER)',/)
      CALL RELMIN(AIC,NV1,DEC,MIN1,MIN2,4H AIC)
      MIN1=MIN1-1
      MIN2=MIN2-1
      WRITE(6,50) MIN1,MIN2
   50 FORMAT(/,10X,'BEST ORDER BY AIC =',I3,/,10X,'2ND BEST ORDER ',
     +'BY AIC =',I3,/)
C
C   COMPUTE ENTROPY MEASURE FOR EACH ORDER
C
C
      LOC=1
      DO 180 I=1,3
      K=MENT(I)
      IF(K.EQ.0) GO TO 180
      K1=NVAR(K)
      DO 170 KK=1,K
      IND(LOC)=INDV(K1)
      COF(LOC)=BEST(K1)
      K1=K1+1
      LOC=LOC+1
```

```
170   CONTINUE
180   CONTINUE
      RETURN
      END
```

APPENDIX C

C.1 Subprogram CSQREG

```
      SUBROUTINE CSQREG(A,NDIM,NIV,IORD,BEST,INDV,MDIM,RVAR,NVAR,NIVN)
C*****************************************************************
C
C    SUBPROGRAM TO PERFORM SEQUENTIAL REGRESSION USING COVARIANCE
C    OR CORRELATION MATRIX A(NIV+1,NIV+1).
C
C    INPUT: A - COVARIANCE MATRIX (COMPLEX)
C           NDIM - ROW DIMENSION OF A IN CALLING PROGRAM
C           NIV - NUMBER OF INDEPENDENT VARIABLES
C           IORD - INTEGER VECTOR CONTAINING INDICES OF VARIABLES
C                  IN THE ORDER THEY ARE TO BE ENTERED INTO THE MODEL
C           MDIM - DIMENSION OF BEST IN CALLING PROGRAM
C
C    OUTPUT: A - SWEPT COVARIANCE MATRIX
C            BEST,INDV - VECTORS OF SUBSET INFORMATION
C                 BEST CONTAINS LEAST SQUARES PARAMETER ESTIMATES
C                 INDV CONTAINS VARIABLE INDICES
C            RVAR - VECTOR OF RESIDUAL VARIANCES
C            IORD - VECTOR CONTAINING INDICES OF VARIABLES IN ORDER
C                   THAT THEY WERE ENTERED WITH VALUES CAUSING
C                   SINGULARITIES IN A OMITTED
C            NIVN - NUMBER OF INDEPENDENT VARIABLES INCLUDED IN
C                   ANALYSIS
C
C    SUBPROGRAMS CALLED: CSWEEP
C
C*****************************************************************
C
      COMPLEX A(NDIM,NDIM),BEST(MDIM)
      DIMENSION INDV(MDIM),IORD(NIV),RVAR(NIV),NVAR(NIV)
      DATA TOL/1.E-20/
      NV=NIV+1
      NIVN=NIV
      VAR=REAL(A(NV,NV))
      LOC=1
      LC2=1
      KOUNT=1
      K=1
   10 ID=IORD(K)
      KOUNT=KOUNT+1
      TEST=REAL(A(ID,ID))**2+AIMAG(A(ID,ID))**2
      IF(TEST.LE.TOL) GO TO 40
      CALL CSWEEP(A,NDIM,NV,ID,ID)
      RVAR(K)=REAL(A(NV,NV))/VAR
```

```
      DO 30 KK=1,K
      KID=IORD(KK)
      IF(KK.NE.1) GO TO 20
      NVAR(LC2)=LOC
      LC2=LC2+1
   20 INDV(LOC)=KID
      BEST(LOC)=-A(NV,KID)
      LOC=LOC+1
   30 CONTINUE
      GO TO 60
   40 NIVN=NIVN-1
      DO 50 I=K,NIVN
      IORD(I)=IORD(I+1)
   50 CONTINUE
      GO TO 10
   60 K=K+1
      IF(KOUNT.LE.NIV) GO TO 10
      RETURN
      END
```

C.2 Subprogram CSWEEP

```
      SUBROUTINE CSWEEP(A,NDIM,N,K1,K2)
C*****************************************************************
C
C   SUBROUTINE TO SWEEP THE NXN COMPLEX MATRIX A ON ITS K1
C   THRU K2 DIAGONAL ELEMENTS (SWP(K)SWP(K)A=A)
C
C   INPUT :
C           A,N,K1,K2
C           NDIM : ROW DIMENSION OF A IN CALLING PROGRAM
C
C   OUTPUT :
C           A
C
C   SUBROUTINES CALLED : NONE
C
C*****************************************************************
C
      COMPLEX D,A(NDIM,NDIM)
      DATA NOUT/6/
C
C   FIX DIAGONAL K :
C
      DO 50 K=K1,K2
C
C   CHECK FOR ZERO :
C
```

```
      TEST=REAL(A(K,K))**2+AIMAG(A(K,K))**2
      IF(TEST.LT.1.E-25) GO TO 99
      D=1./A(K,K)
      A(K,K)=1.
C
C   KTH ROW :
C
      DO 10 I=1,N
   10 A(K,I)=D*A(K,I)
C
C   KTH COLUMN :
C
         DO 20 J=1,N
         IF(J.EQ.K) GO TO 20
         A(J,K)=-A(J,K)*D
   20    CONTINUE
C
C   OTHERS :
C
         DO 40 J=1,N
         IF(J.EQ.K) GO TO 40
            DO 30 I=1,N
            IF(I.EQ.K) GO TO 30
            A(J,I)=A(J,I)+A(J,K)*A(K,I)/D
   30       CONTINUE
   40    CONTINUE
C
C
   50 CONTINUE
C
C
      GO TO 110
   99 WRITE(NOUT,100) K,K1,K2
  100 FORMAT(10X,12,15HTH DIAG OF FROM,1X,12,1X,2HTO,1X,
     112,1X,17HIS ZERO IN CSWEEP)
  110 RETURN
      END
```

C.3 Subprogram AUTDEN

```
      SUBROUTINE AUTDEN(W,N,IDQH,IPLT2,MORD,ALPH,RVARW,SIGO,NVW,
     +ISORT,WLAB)
C*********************************************************************
C
C   THIS SUBPROGRAM COMPUTES A SMOOTHED DENSITY QUANTILE
C   FUNCTION BASED ON THE AUTOREGRESSIVE METHOD OF PARZEN(1979).
C   THIS ROUTINE IS BASED ONE THE ONESAM PROGRAM DENSITY ESTIMATION
C   ROUTINE AND USES MANY OF THE SUBPROGRAMS OF ONESAM. SEE
```

```
C    PARZEN AND ANDERSON (1980) FOR DOCUMENTATION.
C
C    INPUT: W - RAW DATA
C           N - SAMPLE SIZE
C           IDQH - INDICATOR FOR NULL DIST. OF W
C           ISORT - 0 IF W AND RANKW SORTED, 1 OTHERWISE.
C           MORD - MAXIMUM ALLOWABLE ORDER (<=6)
C           IPLT2 - 0--> NO PLOTS
C                   1--> PLOT OF AR DENSITY-QUANTILE FUNCTION
C           WLAB - VARIABLE NAME FOR W IN A4 FORMAT
C
C    OUTPUT: NVW - ORDER OF AUTOREGRESSIVE DENSITY ESTIMATOR
C            ALPH - COEFFICIENTS FOR AUTOREGRESSIVE REPRESENTATION
C            RVARW - RESIDUAL VARIANCE FOR BEST ORDER
C            SIGO - INTEGRATING FACTOR (SIGMA-TILDA FOR NULL MODEL)
C
C    SUBPROGRAMS CALLED: ORD,QHLIN,QTOFQ,WSPACE,FORIER,AUTORG,PARZ,
C                        AREST,FQFNC,MDNRIS,QFIND,PLOTXY,FTERP,MINMAX,
C                        MIN,MAX
C
C*********************************************************************
      DIMENSION W(N),RVAR(5),U(500),QN(500),QL(500),FQ(500),
     +WXS(500),CWXS(500),ILOC(5),T(500),CAT(5),WK1(500),WK2(500)
      DIMENSION CAPT(20)
      COMPLEX A(5),PHI(5),ALPH(5),ALPHA(15),RESVAR
      DATA CAPT/4HUNIV,4HARIA,4HTE D,4HENSI,4HTY-Q,4HUANT,4HILE ,
     +4HFOR ,4HRAND,4HOM V,4HARIA,4HBLE ,8*4H    /
      CAPT(13)=WLAB
      WRITE(6,1) WLAB
    1 FORMAT(//,10X,'UNIVARIATE DENSITY ESTIMATION RESULTS FOR ',
     +'VARIABLE ',A4,//)
      DO 5 I=1,N
      T(I)=W(I)
    5 CONTINUE
      N2=N+2
      M=MORD+1
      IF(M.GT.6) M=6
      MM1=M-1
      H=1./FLOAT(N+1)
      IF(H.LT.0.02) H=0.02
      IF(ISORT.EQ.0) GO TO 10
      CALL ORD(T,N)
   10 CONTINUE
      TMIN=T(1)
C
C    COMPUTE N EQUALLY SPACED U VALUES BETWEEN 0 AND 1
C
      U(1)=0.0
      DO 30 J=1,N
      U(J+1)=FLOAT(J)*H
   30 CONTINUE
C
```

```
C    COMPUTE QUANTILE FUNCTION VIA LINEAR INTERPOLATION
C
       CALL QHLIN(T,N,H,U,QN,0,NQ,TMIN,TINT,N2,WK1,WK2)
C
C    COMPUTE LITTLE Q AND FQ=1/(LITTLE Q)
C
       NP1=NQ+1
       CALL QTOFQ(QN,U,NP1,QL,FQ)
C
C    COMPUTE WEIGHTED SPACINGS (LITTLE D(U)) BASED ON IDQH DIST.
C
       NP1=NQ+1
       CALL WSPACE(WXS,CWXS,NP1,FQ,IDQH,U,SIGO)
C
C    COMPUTE FOURIER TRANSFORM OF WEIGHTED SPACINGS
C
       CALL FORIER(WXS,U(2),N,A,M)
C
C    COMPUTE AUTOREGRESSIVE COEFFICIENTS FOR ORDERS 1 TO M
C
       II=1
       DO 100 K=1,MM1
       KP1=K+1
       CALL AUTORG(A,KP1,M,ALPH,PHI,RESVAR)
       RVAR(K)=REAL(RESVAR)
       ILOC(K)=II
       DO 90 J=1,K
       ALPHA(II)=ALPH(J)
       II=II+1
  90   CONTINUE
 100   CONTINUE
       CALL PARZ(RVAR,M-1,N,CAT,NVW)
       IF(NVW.EQ.0) GO TO 115
       LOC=ILOC(NVW)
       DO 110 I=1,NVW
       ALPH(I)=ALPHA(LOC)
       LOC=LOC+1
 110   CONTINUE
 115   CALL CLPLT1(RVAR,M-1,1,4HRVAR,41,1)
C
C    COMPUTE UNIVARIATE DENSITY-QUANTILE AT 100 POINTS AND PLOT
C
       WRITE(6,120) SIGO
 120   FORMAT(/,10X,'SIGO = ',F10.4)
       RVARW=RVAR(NVW)
       DO 160 I=1,100
       U(I)=FLOAT(I)/101.0
       FI=1.0
       IF(NVW.GT.0) FI=AREST(U(I),NVW,RVARW,ALPH)
       IF(FI.EQ.0.) FI=H
       FQ(I)=FQFNC(U(I),IDQH)/(FI*SIGO)
 160   CONTINUE
```

```
      IF(IPLT2.EQ.1)
     +CALL PLOTXY(U,FQ,100,CAPT,4H   U,4H  FQ,1)
      RETURN
      END
```

C.4 Subprogram LEGP

```
      SUBROUTINE LEGP(N,P,IDEG)
C
C*************************************************************
C
C   SUBROUTINE TO GENERATE 3XN MATRIX P OF COEFFICIENTS
C   OF LEGENDRE POLYNOMIALS. ROW 3 CONTAINS COEFFICIENTS
C   FOR THE LEGENDRE POLYNOMIAL (OVER (-1,1)) OF
C   DEGREE IDEG (DETERMINED BY SUBROUTINE AFTER INITIAL CALL)
C
C   INPUT: N - SAMPLE SIZE (OR HIGHEST ORDER DESIRED)
C          IDEG - 0 --> FOR FIRST CALL
C                 ORDER OF POLYNOMIAL IN THIRD ROW FOR SUBSEQUENT
C                 CALLS (PROVIDED BY ROUTINE)
C
C   OUTPUT: P - THE 3XN MATRIX OF LEGENDRE POLYNOMIAL
C               COEFFICIENTS.
C
C   ALGORITHM: THE SECOND ORDER RECURSION RELATION COMMONLY
C              FOUND IN MOST TEXTBOOKS (SEE, E.G., CHURCHILL,
C              "SPECIAL FUNCTIONS")
C
C*************************************************************
C
      DIMENSION P(3,N)
      IF(IDEG.NE.0) GO TO 30
      DO 20 I=1,3
      DO 10 J=1,N
      P(I,J)=0.0
   10 CONTINUE
   20 CONTINUE
      P(1,1)=1.0
      P(2,2)=1.0
      IDEG=2
      GO TO 70
   30 IDEG=IDEG+1
      DO 60 J=1,IDEG
      P(1,J)=P(2,J)
      P(2,J)=P(3,J)
   60 CONTINUE
   70 CONTINUE
      A=(2.0*FLOAT(IDEG)-1.0)/FLOAT(IDEG)
```

```
      B=FLOAT(IDEG-1)/FLOAT(IDEG)
      II=IDEG+1
      DO 100 J=1,II
      IF(J.GT.1) GO TO 80
      P(3,1)=-B*P(1,1)
      GO TO 100
   80 P(3,J)=A*P(2,J-1)-B*P(1,J)
  100 CONTINUE
      RETURN
      END
```

APPENDIX D

D.1 Subprogram PLOTXY

```
      SUBROUTINE PLOTXY(X,Y,N,CAPT,NAMX,NAMY,IOPT)
C*************************************************************
C
C   SUBROUTINE TO PRINT AND PRINTER PLOT THE N-VECTOR Y AS A
C   FUNCTION OF X.
C
C   INPUT : N,X,Y - X IS ORDERED ON INPUT AND Y(I)=Y(X(I))
C           CAPT - LITERAL CONSTANT FOR TITLE OF PLOT IN 20A4 FORMAT
C           NAMX,NAMY : 4 CHARACTER LITERAL CONSTANTS GIVING
C                       LABELS FOR X AND Y
C           IOPT : 1,2 (POINT OR BAR PLOT)
C
C   SUBROUTINES CALLED : FTERP,MAX,MIN
C
C*************************************************************
C
      DIMENSION X(N),Y(N),T(46),YI(46),CAPT(20),AL(101)
      DATA NOUT/6/
      DATA BLANK,DOT,Z,SL,PLUS/1H ,1H.,1H*,1H|,1H+/
C
      MM=81
      IOPTY=0
      IF(N.GT.19) GO TO 11
      WRITE(NOUT,10) N
   10 FORMAT(10X,'SAMPLE SIZE OF ',I2,' IS TOO SMALL TO PERFORM ',
     +'INTERPOLATION IN PLOTXY.')
      GO TO 100
   11 CONTINUE
      WRITE(NOUT,13) CAPT
   13 FORMAT(1H1,33X,20A4,/)
C
C   CREATE T VECTOR OF EQUALLY SPACED X AND INTERPOLATE TO OBTAIN
C   CORRESPONDING Y VALUES
C
      DEC=(X(N)-X(1))/45.0
      DO 15 I=1,46
      T(I)=X(1)+FLOAT(I-1)*DEC
   15 CONTINUE
      CALL FTERP(X,Y,T,YI,N,46)
C
C   INITIALIZE AL :
C
      ON=(MM-1)/2
      DO 20 J=1,MM
```

```
   20  AL(J)=DOT
       WRITE(NOUT,25)  NAMX,NAMY, (AL(J),J=1,MM)
   25  FORMAT(/,16X,A4,6X,A4/10X,20(1H-),2X,101A1)
       DO 30 J=1,MM
   30  AL(J)=BLANK
       AL(1)=SL
       AL(MM)=SL
C
C    FIND MAX AND MIN :
C
       CALL MAX(Y1,46,YMAX,IND)
       CALL MIN(Y1,46,YMIN,IND)
       RY=1.2*(YMAX-YMIN)
       IF(RY.LT.1.E-20)  IOPTY=1
C
C    PLOT :
C
       DO 40 J=1,46
       IF(IOPTY.EQ.1)  GO TO 36
       C1=(Y1(J)-YMIN)/RY
       C1=2.*(C1-.5)
       GO TO 37
   36  C1=0.
   37  K=ON*(C1+1.)+2.5
       AL(K)=Z
       IF(IOPT.EQ.1)  GO TO 35
       DO 39 I=1,K
   39  AL(I)=Z
   35  CONTINUE
       WRITE(NOUT,38)  T(J),Y1(J), (AL(I),I=1,MM)
   38  FORMAT(10X,F10.4,1X,F9.4,2X,101A1)
       AL(K)=BLANK
       IF(IOPT.EQ.1)  GO TO 40
       DO 41 I=2,K
   41  AL(I)=BLANK
   40  CONTINUE
       DO 50 I=1,MM
   50  AL(I)=DOT
       AL(1)=PLUS
       AL(MM)=PLUS
       WRITE(NOUT,60)  (AL(I),I=1,MM)
   60  FORMAT(10X,20(1H-),2X,101A1)
C
C
       YMAX=RY+YMIN
       WRITE(NOUT,70)  YMIN,YMAX
   70  FORMAT(27X,F10.4,70X,F10.4)
  100  CONTINUE
       RETURN
       END
```

D.2 Subprogram PLTXYZ

```
      SUBROUTINE PLTXYZ(X,Y,Z,N,CAPT,NAMX,NAMY,NAMZ,XISE,XMAXD,XMSE)
C*************************************************************
C
C   SUBROUTINE TO PRINT AND PRINTER PLOT THE N-VECTORS Y AND Z AS
C   A FUNCTION OF X ON THE SAME COORDINATE SYSTEM.
C
C   INPUT : N,X,Y,Z - X IS ORDERED ON INPUT AND Y(I)=Y(X(I)),ETC.
C           CAPT - LITERAL CONSTANT FOR TITLE OF PLOT IN 20A4 FORMAT
C           NAMX,NAMY,NAMZ : 4 CHARACTER LITERAL CONSTANTS GIVING
C                            LABELS FOR X, Y, AND Z
C           XISE,XMAXD,XMSE - SQUARED ERROR DIAGNOSTICS
C
C   SUBROUTINES CALLED : FTERP,MAX,MIN
C
C*************************************************************
C
      DIMENSION X(N),Y(N),Z(N),T(46),YI(46),ZI(46),CAPT(20),AL(101)
      DATA NOUT/6/
      DATA BLANK,DOT,STAR,SL,PLUS/1H ,1H.,1H*,1H|,1H+/
      DATA SM,SO/1HM,1HO/
C
      MM=81
      IOPTY=0
      IF(N.GT.19) GO TO 11
      WRITE(NOUT,10) N
   10 FORMAT(10X,'SAMPLE SIZE OF ',I2,' IS TOO SMALL TO PERFORM ',
     +'INTERPOLATION IN PLTXYZ.')
      GO TO 100
   11 CONTINUE
      WRITE(6,13) XISE,XMAXD,XMSE
   13 FORMAT(1H1,9X,85(1H-),/10X,'|  INTEGRATED SQUARE ERROR =',E10.4,
     +5X,'MAXIMUM ABSOLUTE DIFFERENCE =',E10.4,'  |',/,10X,'|  MEAN ',
     +'SQUARE ERROR =',E10.4,52X,'|',/,10X,85(1H-))
C
C   CREATE T VECTOR OF EQUALLY SPACED X AND INTERPOLATE TO OBTAIN
C   CORRESPONDING Y AND Z VALUES
C
      DEC=(X(N)-X(1))/45.0
      DO 15 I=1,46
      T(I)=X(1)+FLOAT(I-1)*DEC
   15 CONTINUE
      CALL FTERP(X,Y,T,YI,N,46)
      CALL FTERP(X,Z,T,ZI,N,46)
C
C   INITIALIZE AL :
C
      ON=(MM-1)/2
      DO 20 J=1,MM
```

```
   20  AL(J)=DOT
       WRITE(NOUT,25) NAMX,NAMY,NAMZ,(AL(J),J=1,MM)
   25  FORMAT(/,16X,A4,4X,2H*-,A4,4X,2HO-,A4/10X,30(1H-),2X,101A1)
       DO 30 J=1,MM
   30  AL(J)=BLANK
       AL(1)=SL
       AL(MM)=SL
C
C    FIND MAX AND MIN :
C
       CALL MAX(Y1,46,YMAX,IND)
       CALL MIN(Y1,46,YMIN,IND)
       CALL MAX(Z1,46,ZMAX,IND)
       CALL MIN(Z1,46,ZMIN,IND)
       IF(ZMIN.LT.YMIN) YMIN=ZMIN
       IF(ZMAX.GT.YMAX) YMAX=ZMAX
       RY=1.2*(YMAX-YMIN)
       IF(RY.LT.1.E-20) IOPTY=1
C
C    PLOT :
C
       DO 40 J=1,46
       IF(IOPTY.EQ.1) GO TO 36
       C1=(Y1(J)-YMIN)/RY
       C1=2.*(C1-.5)
       C2=(Z1(J)-YMIN)/RY
       C2=2.*(C2-.5)
       GO TO 37
   36  C1=0.
       C2=0.
   37  KY=ON*(C1+1.)+2.5
       KZ=ON*(C2+1.)+2.5
       AL(KY)=STAR
       AL(KZ)=SO
       IF(KY.EQ.KZ) AL(KY)=SM
       WRITE(NOUT,38) T(J),Y1(J),Z1(J),(AL(I),I=1,MM)
   38  FORMAT(10X,F10.4,1X,F9.4,1X,F9.4,2X,101A1)
       AL(KY)=BLANK
       AL(KZ)=BLANK
   40  CONTINUE
       DO 50 I=1,MM
   50  AL(I)=DOT
       AL(1)=PLUS
       AL(MM)=PLUS
       WRITE(NOUT,60)  (AL(I),I=1,MM)
   60  FORMAT(10X,30(1H-),2X,101A1)
C
C
       YMAX=RY+YMIN
       WRITE(NOUT,70) YMIN,YMAX
   70  FORMAT(37X,F10.4,70X,F10.4)
       WRITE(NOUT,80) CAPT
```

```
   80  FORMAT(/,10X,20A4,/)
  100  CONTINUE
       RETURN
       END
```

D.3 Subprogram FTERP

```
       SUBROUTINE FTERP(U,V,X,F,N,M)
C*************************************************************
C
C    SUBROUTINE TO PERFORM LINEAR INTERPOLATION ON V TO
C    OBTAIN F AT THE M X VALUES
C
C    INPUT: U - VECTOR OF VALUES AT WHICH V EVALUATED
C           V - FUNCTION VALUES TO INTERPOLATE
C           X - VALUES AT WHICH INTERPOLATED FUNCTION TO BE
C               EVALUATED
C           N - DIMENSION OF VECTORS U AND V
C           M - DIMENSION OF VECTORS X AND F
C
C    NOTE: ALL ABSCISSA VECTORS MUST BE ORDERED
C
C    OUTPUT: F - INTERPOLATED FUNCTION VALUES
C
C*************************************************************
       DIMENSION U(N),V(N),X(M),F(M)
       IF(N.EQ.M) GO TO 100
       II=1
       DO 60 I=1,M
   10  IF(X(I)-U(II))20,40,50
   20  IF(II.NE.1) GO TO 30
       F(I)=V(1)+(V(2)-V(1))*(X(I)-U(1))/(U(2)-U(1))
       GO TO 60
   30  F(I)=V(II-1)+(V(II)-V(II-1))*(X(I)-U(II-1))/(U(II)-U(II-1))
       GO TO 60
   40  F(I)=V(II)
       GO TO 60
   50  II=II+1
       IF(II.LT.N) GO TO 10
       II=N
       GO TO 30
   60  CONTINUE
  100  RETURN
       END
```

D.4 Subprogram SEDIAG

```
      SUBROUTINE SEDIAG(F,G,N,RANGE,XISE,XMAXD,XMSE)
C*****************************************************************
C
C    SUBROUTINE TO COMPUTE VARIOUS SQUARED ERROR DIAGNOSTICS
C    BY 'RAW' NUMERICAL INTEGRATION (RIEMANN SUMS).
C
C    INPUT: F,G - VECTORS CONTAINING FUNCTION VALUES CORRESPONDING
C                 TO THE SAME ARGUMENT, I.E., F(I)=F(X(I))
C                 CORRESPONDS TO G(I)=G(X(I)). F AND G MUST HAVE
C                 BEEN EVALUATED AT EQUALLY SPACED X VALUES.
C           N - DIMENSION OF F AND G
C           RANGE - RANGE OF X VALUES OVER WHICH F AND G ARE
C                   COMPUTED.
C
C    OUTPUT: XISE - INTEGRATED SQUARED ERROR
C            XMAXD - MAXIMUM ABSOLUTE DIFFERENCE BETWEEN F AND G
C            XMSE - MEAN SQUARED ERROR ASSUMING EXPECTATION TAKEN
C                   WITH RESPECT TO F.
C
C*****************************************************************
      DIMENSION F(N),G(N)
      XMAXD=0.0
      XISE=0.0
      XMSE=0.0
      DO 10 I=1,N
      DIF=(F(I)-G(I))*(F(I)-G(I))
      IF(XMAXD.LT.DIF) XMAXD=DIF
      XISE=XISE+DIF
      XMSE=XMSE+DIF*F(I)
   10 CONTINUE
      XMAXD=SQRT(XMAXD)
      XISE=XISE*RANGE/FLOAT(N)
      XMSE=XMSE*RANGE/FLOAT(N)
      RETURN
      END
```

ABBREVIATION INDEX

AUTHOR INDEX


ND
ATE